Bayesian Brittleness: Why no Bayesian model is "good enough"

Houman Owhadi, Clint Scovel, Tim Sullivan

April 26, 2013

Abstract

Although it is known that Bayesian estimators may fail to converge or may converge towards the wrong answer (i.e. be inconsistent) if the probability space is not finite or if the model is misspecified (i.e. the data-generating distribution does not belong to the family parametrized by the model), it is also a popular belief that a "good" or "close" enough model should have good convergence properties. This paper incorporates Bayesian priors into the Optimal Uncertainty Quantification (OUQ) framework [86] and in doing so reveals extreme brittleness in Bayesian inference. These brittleness results demonstrate that, contrary to popular belief, there is no such thing as a "close enough" model in Bayesian inference in the following sense: we derive optimal lower and upper bounds on posterior values obtained from models that exactly capture an arbitrarily large (but finite) number of finite-dimensional marginals of the data-generating distribution and/or that are arbitrarily close to the data-generating distribution in the Prokhorov or total variation metrics; these bounds show that such models may still make the largest possible prediction error after conditioning on an arbitrarily large number of sample data. Therefore, under model misspecification, and without stronger assumptions than (arbitrary) closeness in Prokhorov or total variation metrics, Bayesian inference offers no better guarantee of accuracy than arbitrarily picking a value between the essential infimum and supremum of the quantity of interest. In particular, an unscrupulous practitioner could slightly perturb a given prior and model to achieve any desired posterior conclusions.

Finally, this paper also addresses the non-trivial technical questions of how to incorporate priors in the OUQ framework. In particular, we develop the necessary measure theoretical foundations in the context of Polish spaces, so that simultaneously prior measures can be put on subsets of a product space of functions and measures and important quantities of interest are measurable. We also develop the reduction theory for optimization problems over measures on product spaces of measures and functions, thus laying down the foundations for the scientific computation of optimal statistical estimators.

2010 Mathematics Subject Classification: 62A01, 62E20, 62F12, 62F15, 62G20, 62G35.
Keywords: Bayesian inference, misspecification, robustness, uncertainty quantification, optimal uncertainty quantification.
Houman Owhadi: Corresponding author, California Institute of Technology, owhadi@caltech.edu
Clint Scovel: California Institute of Technology, clintscovel@gmail.com
Tim Sullivan: University of Warwick, Tim.Sullivan@warwick.ac.uk

Contents

1 Introduction

Throughout science and industry, Bayesian methods are increasingly popular tools for the understanding of uncertainty in often complicated contexts, and they impact the making of sometimes critical decisions. It is probably fair to say that, despite their popularity and documented successes, Bayesian methods have always attracted some degree of controversy and opposition: see e.g. [59] and rejoinders for a recent academic discussion, and [78, 81] for less formal treatments. Often, this opposition is philosophical in nature, particularly with regard to the subjective interpretation of the probabilities involved, which is something that remains counter-intuitive to many commentators: see [50, par. 35 & 37] for a recent example in law. However, there are also analytical reasons to be wary about the application of Bayesian methods: there is now half a century's

worth of examples of situations in which the Bayesian posterior behaves in apparently perverse ways and yields predictions that are, by any objective measure, wrong.

It is, in fact, now well understood that Bayesian methods may fail to converge or may converge towards the wrong solution if the underlying probability mechanism allows an infinite number of possible outcomes [42] and that, in these non-finite-probability-space situations, this lack of convergence (commonly referred to as *Bayesian inconsistency*) is the rule rather than the exception [43]. Conversely, it is known from the Bernstein–von Mises Theorem [24, 76, 110] that consistency (convergence of the Bayesian posterior to the data-generating distribution in the limit of observing infinite amounts of sample data) does indeed hold, under some regularity conditions, if the data-generating distribution belongs to the finite-dimensional family of distributions parametrized by the model (i.e. if the model is *well specified*).

However, although it is known that this convergence may fail under *model misspeciation* [119, 61, 88, 1, 2, 74, 77, 62] (i.e. when the data-generating distribution does not belong to the finite-dimensional family of distributions parametrized by the model, as illustrated in Figure 2.1) it is also a popular belief that a "close enough" (or "good enough") model should have good convergence properties: see e.g. [46, 95, 47]. This belief echoes G. E. P. Box's statement [30, p. 424] that "essentially, all models are wrong, but some are useful" and question [30, p. 74] "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful?"

The brittleness results of this paper (Theorems 5.12, 6.4 and 6.10) show that, contrary to this popular belief, there is no such thing as a "close enough" model in Bayesian inference in the following sense: suppose that one calculates optimal bounds (i.e. least upper and greatest lower bounds) on posterior values with respect to a Bayesian model that exactly captures an arbitrarily large (but finite) number of finite-dimensional marginals of the data-generating distribution and/or that are arbitrarily close to the true (data generating) distribution in the Prokhorov or total variation metrics; our results show that such models may still make the largest possible prediction error even after conditioning on an arbitrarily large number of sample data observations, and also in the limit as the number of observations tends to infinity. Therefore, these brittleness theorems suggest that, under model misspecification and without stronger assumptions than closeness in the Prokhorov and/or total variation metrics, Bayesian inference offers no better guarantee of accuracy than arbitrarily picking a value between the essential infimum and supremum of the quantity of interest. In particular, an unscrupulous practitioner can slightly perturb a given prior and model to achieve any desired posterior conclusions.

As noted by G. E. P. Box, for complex systems, all models are misspecified. Indeed, for such systems, the data-generating distribution is a point in an infinite dimensional space of measures whereas Bayesian models, in their common (parametric) applications, form finite-dimensional subspaces of these infinite dimensional spaces of measures. This view, combined with our brittleness results, establishes that, in complex systems, although Bayesian methods may work well, they may also work very poorly. In either case, without more information, one will not know.

1.1 Structure of the paper and main results

This paper is structured as follows: Section 2 reviews questions of Bayesian consistency, inconsistency, model misspecification, and robustness through a motivating analysis. Section 3 incorporates Bayesian priors into the Optimal Uncertainty Quantification (OUQ) framework [86]. In the OUQ framework, Uncertainty Quantification (UQ) is formulated as an optimization problem (over an infinite-dimensional set of functions and measures) corresponding to extremizing (i.e. finding worst and best case scenarios) probabilities of failure or other quantities of interest, subject to the constraints imposed by the scenarios compatible with the assumptions and information. In particular, the OUQ framework allows for the treatment of systems of partially-known probability measures and response functions; such systems arise naturally in studies of materials, financial systems, insurance against catastrophes, medicine and law. This generalization of the OUQ framework to Bayesian priors requires the development of measure theoretical foundations so that simultaneously prior measures can be put on subsets of a product space of functions and measures and important quantities of interest are measurable. This non-trivial and highly technical task is achieved in Section 7 in the context of Polish topological spaces.

In this generalization, priors are probability measures on spaces of measures and functions, and computing optimal bounds on prior values (given a set of priors) requires solving problems in which the optimization variables are measures on spaces of measures and functions. Section 4 shows how such optimization problems can, under general conditions, be reduced to the iteration of two optimization problems in which the optimization variables are measures and functions, where then we can apply the reduction theorems of [86].

Our motivation for addressing the measurability and reduction issues raised in sections 7 and 4 goes beyond the investigation of the Brittleness of Bayesian Inference as we also seek to lay down the foundations for the scientific computation of optimal statistical estimators (i.e., using computers to find estimators with minimal statistical errors, this constitutes a sequel work).

Section 5 provides similar reduction theorems for the computation of optimal bounds on posterior values given a set of priors and the observation of the data. These reduction theorems lead to the Brittleness results (Theorems 5.12, 6.4 and 6.10). In particular, Section 6 presents the Brittleness under Local Misspecification theorems (theorems 6.4 and 6.10). That is, given a Bayesian model, Theorem 6.4 provides optimal bounds on posterior values for priors that are at arbitrarily small distance (in the Prokhorov or total variation metrics) from a given Bayesian model. Theorems 6.4 and 6.10 show that these optimal bounds on posterior values are the essential supremum and infimum of the quantity of interest irrespective of the size of data and of the size of the metric neighbourhood around the Bayesian model. Sections 8 and 9 contain the proofs of our results.

1.2 Notation and Conventions

Throughout, for a topological space $\mathcal{Y}$, $\mathcal{B}(\mathcal{Y})$ will denote the Borel σ-algebra of subsets of $\mathcal{Y}$ and $\mathcal{M}(\mathcal{Y})$ will denote the space of Borel probability measures. For an alternative σ-algebra $\Sigma_{\mathcal{Y}}$ of subsets of $\mathcal{Y}$ the set of probability measures on the σ-algebra $\Sigma_{\mathcal{Y}}$ will be denoted $\mathcal{M}(\Sigma_{\mathcal{Y}})$. If $\mathcal{Y}$ is metrizable, $\mathcal{M}(\mathcal{Y})$ is endowed with the weak-$*$ topology and the corresponding Borel σ-algebra unless specified otherwise. For a mapping between topological spaces, the term measurable will mean Borel measurable unless specified otherwise. Moreover, suprema over the empty set will have the value $-\infty$ and infimima over the empty set the value $+\infty$.

2 Bayesian Inconsistency and Model Misspecification: a motivating analysis

To motivate the results of this paper, this section will analyse and review questions of Bayesian consistency, inconsistency, model misspecification, and robustness. There is, of course, a large literature on these topics, and we will not attempt to be exhaustive in providing references; rather, our aims are: first, to give a short reminder on how Bayesian inference is currently employed in Uncertainty Quantification (UQ); second, to identify issues and popular beliefs about what one actually learns from Bayesian inference, and thereby motivate the results of this paper; and, last, to present sufficient references that the interested reader can find technical justification for the formal manipulations of this section.

In this section, we are interested in estimating

$$\Phi(\mu^\dagger) \tag{2.1}$$

where Φ is a known *quantity of interest* function and $\mu^\dagger$ is an unknown (or partially known) probability measure on $\mathcal{X}$. For the purposes of exposition, in this section, we assume that $\mathcal{X} = \mathbb{R}^k$. One example of a quantity of interest, when $\mathcal{X} = \mathbb{R}$, is $\Phi(\mu^\dagger) := \mu^\dagger[X \geq a]$ (the probability that the random variable X distributed according to $\mu^\dagger$ exceeds the threshold value a). We also assume that we are given n independent samples $d_1, \ldots, d_n$, each distributed according to $\mu^\dagger$.

We will now present the parametric Bayesian answer to this problem. For the purposes of exposition, in this section, we restrict our attention to parametric Bayesian inference. We first introduce $\{\mu(\,\cdot\,,\theta)\}_{\theta\in\Theta}$ a family of probability distributions on $\mathcal{X}$ parameterized by $\theta \in \Theta$ (and commonly referred to as the *model class*). For the sake of simplicity here we also assume that $\Theta = \mathbb{R}^\ell$. Let

$$\mathcal{A}_0 := \big\{\mu(\,\cdot\,,\theta) \,\big|\, \theta \in \Theta\big\}. \tag{2.2}$$

Note that $\mathcal{A}_0$ is a subset of $\mathcal{M}(\mathcal{X})$ that may or may not contain $\mu^\dagger$. If $\mu^\dagger \notin \mathcal{A}_0$, then the model is said to be *misspecified*; otherwise, the model is said to be *well specified*.

We next introduce $p_0 \in \mathcal{M}(\Theta)$, a probability distribution on Θ (the *prior distribution* on θ). Let π_0 be the push-forward of p_0 under the map $\theta \mapsto \mu(\,\cdot\,,\theta)$ and observe that π_0

is a probability distribution on $\mathcal{A}_0$, i.e. $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$, and that π_0 is the distribution of the random measure $\mu(\,\cdot\,,\theta)$ when θ is distributed according to p_0.

The next step is then to estimate $\Phi(\mu^\dagger)$ via conditioning. Let $p_n \in \mathcal{M}(\Theta)$ be the posterior distribution of θ given the observation of the i.i.d. samples $d_1, \ldots, d_n$, as obtained using Bayes' formula, and let π_n be the push-forward of p_n. The Bayesian estimate of $\Phi(\mu^\dagger)$ is therefore

$$\mathbb{E}_{\mu\sim\pi_n}[\Phi(\mu)]. \tag{2.3}$$

For the purposes of exposition, we assume that the measures $\mu(\,\cdot\,,\theta)$ and $\mu^\dagger$ are all absolutely continuous with respect to the Lebesgue measure and write $\beta(\,\cdot\,,\theta)$ and $\beta^\dagger$ for their densities, which we assume to be continuous. Similarly, we assume that the measure p_0 is absolutely continuous with respect to the Lebesgue measure and, abusing notation, write p_0 for both the measure p_0 and its (continuous) density, and similarly for $p_n(\,\cdot\,)$, the posterior density of θ on Θ given the observation the samples $d_1, \ldots, d_n$. We will now examine the convergence properties of the sequence of posterior densities $p_n(\theta)$ as $n \to \infty$. This analysis being classical (see for instance [84] and references therein), our purpose is not to provide rigorous justifications but rather to familiarize the reader with the mechanisms regarding the convergence of posteriors.

We have

$$p_n(\theta) = \frac{p_0(\theta)\prod_{j=1}^n \beta(d_j,\theta)}{\int_\Theta p_0(\theta')\prod_{j=1}^n \beta(d_j,\theta')\,\mathrm{d}\theta'} \equiv \frac{p_0(\theta)\prod_{j=1}^n \beta(d_j,\theta)}{\mathbb{E}_{p_0}[\prod_{j=1}^n \beta(d_j,\,\cdot\,)]} \tag{2.4}$$

which we write as

$$p_n(\theta) = \frac{p_0(\theta)e^{nL_n(\theta)}}{\int_\Theta p_0(\theta')e^{nL_n(\theta')}\,\mathrm{d}\theta'} \equiv \frac{p_0(\theta)e^{nL_n(\theta)}}{\mathbb{E}_{p_0}[e^{nL_n(\,\cdot\,)}]}, \tag{2.5}$$

where

$$L_n(\theta) := \frac{1}{n}\sum_{j=1}^n \log\beta(d_j,\theta). \tag{2.6}$$

Recall that $\prod_{j=1}^n \beta(d_j,\theta)$ is commonly known as the *likelihood* and $L_n(\theta)$ as the (*sample*) *average log-likelihood*.

Consistency and the large-sample limit. Now observe that if $\log\beta(d_j,\theta)$ is integrable then it follows from the Law of Large Numbers that $L_n(\theta)$ converges almost surely, as $n \to \infty$, to the *expected log-likelihood* $L(\theta)$ defined by

$$L(\theta) := \int_{\mathcal{X}} \beta^\dagger(x)\log\big(\beta(x,\theta)\big)\,\mathrm{d}x\,. \tag{2.7}$$

Assuming that $L(\theta)$ has a unique maximizer $\theta^* \in \Theta$ — known as the *maximum likelihood estimator* (MLE) — and that p_0 is strictly positive in every neighborhood of θ^*, it follows under assumptions on the regularity of β (or local strict convexity in the neighborhood of θ^*) that $p_n(\theta)$ converges, almost surely, as $n \to \infty$, towards a Dirac mass supported

at θ^*. Therefore, assuming Φ to be sufficiently regular, the Bayesian posterior estimate of $\Phi(\mu^\dagger)$, i.e.,

$$\int_\Theta \Phi\big(\mu(\,\cdot\,,\theta)\big)p_n(\theta)\,\mathrm{d}\theta \tag{2.8}$$

converges almost surely as $n \to \infty$ to

$$\Phi\big(\mu(\,\cdot\,,\theta^*)\big) \tag{2.9}$$

Note that

$$L(\theta) = \mathrm{Ent}(\beta^\dagger) - D_{\mathrm{KL}}\big(\beta^\dagger \big\| \beta(\,\cdot\,,\theta)\big),$$

where $\mathrm{Ent}(\beta^\dagger) := -\int_{\mathcal{X}} \beta^\dagger(x)\log\beta^\dagger(x)\,\mathrm{d}x$ is the *entropy* of $\beta^\dagger$ and D_{KL} denotes the *Kullback–Leibler divergence* defined by

$$D_{\mathrm{KL}}\big(\beta^\dagger \big\| \beta(\,\cdot\,,\theta)\big) := \mathbb{E}_{x\sim\beta^\dagger}\left[\log\frac{\beta^\dagger(x)}{\beta(x,\theta)}\right]. \tag{2.10}$$

It follows that θ^* is also the minimizer of $D_{\mathrm{KL}}\big(\beta^\dagger \big\| \beta(\,\cdot\,,\theta)\big)$ with respect to θ, i.e. the MLE θ^* is characterized by the property that $\mu(\,\cdot\,,\theta^*)$ is the distribution having minimal relative entropy to $\mu^\dagger$ in the model class $\{\mu(\,\cdot\,,\theta)\}_{\theta\in\Theta}$.

An immediate consequence of this observation is the fact if the model is not misspecified, i.e. if $\mu^\dagger$ is an element $\mu(\,\cdot\,,\theta^\dagger)$ of the model class, then $\theta^* = \theta^\dagger$, $\mu(\,\cdot\,,\theta^*) = \mu^\dagger$, and the Bayesian estimate (2.8) is asymptotically exact in the limit as $n \to \infty$. In this situation, the Bayesian estimate is said to be *consistent*.

This convergence result is known as the Bernstein–von Mises Theorem (see for instance [84, Theorem 5]) or as the Bayesian Central Limit Theorem, since the limiting posterior can even be described in a more refined way as being asymptotically normal and not just a point mass. The condition that every open neighbourhood of $\theta^\dagger$ has strictly positive p_0-probability is known informally as Cromwell's Rule[1].

What happens when the model is misspecified? To provide an illustrative answer to this question, consider the family of Gaussian models $\{\beta(\,\cdot\,,\theta) \mid \theta = (c,\sigma) \in \mathbb{R}\times\mathbb{R}_+\}$, where

$$\beta(x,c,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-c)^2}{2\sigma^2}\right).$$

What will happen when this model is exposed to data coming from a potentially non-Gaussian truth $\mu^\dagger$, with density $\beta^\dagger$, that has a well-defined mean $c^\dagger$ and standard deviation $\sigma^\dagger$? By the above considerations, θ^* maximizes the expected log-likelihood (2.7) with respect to θ, and the expected log-likelihhod is simply

$$L(\theta) = -\int_{\mathbb{R}} \beta^\dagger(x)\frac{(x-c)^2}{2\sigma^2}\,\mathrm{d}x - (\log\sigma)\int_{\mathbb{R}}\beta^\dagger(x)\,\mathrm{d}x - \log\sqrt{2\pi}\,. \tag{2.11}$$

[1]Since the posterior cannot possibly concentrate on a point outside the support of the prior, having a globally-supported prior and hence not ruling out a priori any $\theta \in \Theta$ as a possible $\theta^\dagger$ can be seen as a Bayesian version of Oliver Cromwell's famous injunction to the Synod of the Church of Scotland in 1650: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

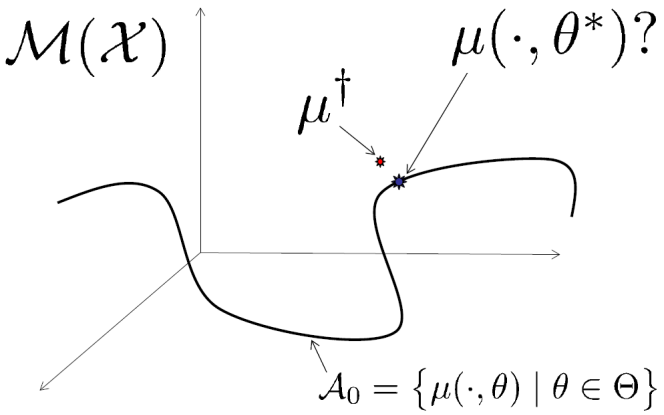

Figure 2.1: According to the brittleness theorems 5.12, 6.4 and 6.10, the Bayesian model $\{\mu(\,\cdot\,, \theta)\}_{\theta \in \Theta}$ may be arbitrarily close to the (true) data generating distribution $\mu^\dagger$ (in Prokhorov and/or total variation metrics or in terms of the number of finite-dimensional marginals of the data generating distribution that are exactly captured) and still make the largest possible prediction error after conditioning on an arbitrary large number of sample data. Note that, for complex systems systems, the data generating distribution is a point in the infinite dimensional space of measures $\mathcal{M}(\mathcal{X})$ whereas the Bayesian model is a finite-dimensional subspace of $\mathcal{M}(\mathcal{X})$. Note also that if the truth $\mu^\dagger$ and the model class are not mutually absolutely continuous, then the relative entropy distance from $\mu^\dagger$ to the model class will be infinite, and this is generic in infinite-dimensional / non-parametric settings.

A quick calculation using partial derivatives shows that $\theta^* = (c^*, \sigma^*)$ maximizes (2.11) if and only if $c^* = c^\dagger$ and $\sigma^* = \sigma^\dagger$. That is, the Bayesian estimate (2.3) of $\Phi(\mu^\dagger)$, for *any* distribution $\mu^\dagger$ of mean $c^\dagger$ and standard deviation $\sigma^\dagger$, converges almost surely as the number of sample data goes to infinity, towards $\Phi\big(\mu(\,\cdot\,, (c^\dagger, \sigma^\dagger))\big)$, where $\mu(\,\cdot\,, (c^\dagger, \sigma^\dagger))$ is the unique Gaussian distribution on $\mathbb{R}$ with mean $c^\dagger$ and standard deviation $\sigma^\dagger$.

However, now there is a problem: there are many different probability distributions μ on $\mathbb{R}$ that have the same first and second moments as $\mu^\dagger$ but have, say, different higher-order moments, or different quantiles. Predictions of those other moments or quantiles using $\mu(\,\cdot\,, (c^\dagger, \sigma^\dagger))$ can be inaccurate by orders of magnitude. A trivial, albeit extreme, example is furnished by $\Phi(\mu) := \mathbb{E}_\mu\big[|X - c_\mu| \geq t\sigma_\mu\big]$ (where c_μ and σ_μ denote the mean and standard deviation of μ). Under the Gaussian model,

$$\mathbb{P}\big[|X - c_\mu| \geq t\sigma_\mu\big] = 1 + \operatorname{erf}\left(-\frac{t}{\sqrt{2}}\right),$$

whereas the extreme cases that prove the sharpness of Chebyshev's (Markov's optimal) inequality have

$$\mathbb{P}\big[|X - c_\mu| \geq t\sigma_\mu\big] = \min\left\{1, \frac{1}{t^2}\right\}.$$

In the case of the archetypically rare "6σ event", the ratio between the two is approximately 1.4×10^7. This comparison is, of course, an almost perversely extreme comparison: it would be obvious to any observer with only moderate amounts of sample data that the data were being drawn from a highly non-Gaussian distribution. However, it is not inconceivable that the true distribution $\mu^\dagger$ has a Gaussian-looking bulk but tails that are significantly fatter than those of a Gaussian, and the difference may be difficult to establish using reasonable amounts of sample data; yet, it is those tails that drive the occurrence of "Black Swans", catastrophically high-impact but low-probability outcomes. The results of this paper show that this situation is generic, and cannot be avoided no matter how many moments or integrals of arbitrary test functions of the truth $\mu^\dagger$ are matched nor how "close" $\mu^\dagger$ is to the class $\{\mu(\,\cdot\,, \theta)\}_{\theta \in \Theta}$.

2.1 Bayesian Inconsistency and Model Misspecification

To quote [84], "[w]hile for a Bayesian statisfician the analysis ends in a certain sense with the posterior, one can ask interesting questions about the the properties of posterior-based inference from a frequentist point of view." Many of these questions are asymptotic in nature: for example, in the limit of infinitely many independent $\mu^\dagger$-distributed samples, will the posterior converge in a suitable sense to $\mu^\dagger$ regardless of the initial choice of prior π? This property is referred to as *consistency*[2]; a general survey of consistency results is found in [113]. As noted above, the consistency theorem is generically known

[2]Sometimes the term *frequentist consistency* is used, reflecting the fact that it lies outside the strict Bayesian worldview, and, to a fundamentalist Bayesian, even to say that the data are generated by $\mu^\dagger$ is a frequentist heresy.

as the Bernstein–von Mises theorem [24, 110], although the earliest rigorous proofs are due to Doob [44] and Le Cam [76].

Unfortunately, Cromwell's Rule is only necessary, and not sufficient, to ensure consistency. In fact, consistency is far from being a generic property, and once the probability space contains infinitely many points (and hence any parameter space Θ that parametrizes all probability measures on that probability space is infinite-dimensional), inconsistency is not the exception, but the rule [43]. In [54, Sec. 5], Freedman considered a countable index set $\mathbb{N} := \{1, 2, \dots\}$ and the parameter space

$$\Theta := \left\{ \theta \colon \mathbb{N} \to [0,1] \,\middle|\, \sum_{i \in \mathbb{N}} \theta(i) = 1 \right\}.$$

Each θ gives rise to a probability distribution $\mathbb{P}_\theta = \mu(\,\cdot\,, \theta)$ under which the observations $X_1, X_2, \dots$ are IID with $\mathbb{P}_\theta[X_n = i] = \theta(i)$. The problem is assumed to be well-specified, so that one particular $\theta^\dagger \in \Theta$ is considered to be the "true" parameter value, and the frequentist data-generating distribution is $\mu^\dagger = \mathbb{P}_{\theta^\dagger} = \mu(\,\cdot\,, \theta^\dagger)$. Theorem 5 of [54] shows that, when $\operatorname{supp}(\mu^\dagger)$ is infinite, given any "spurious" probability distribution $\mathbb{Q} = \mathbb{P}_q$, there exists a prior probability measure π on Θ that has $\theta^\dagger$ in its support, such that the posterior of π $\mu^\dagger$-a.s. concentrates on q in the limit of observing infinitely many i.i.d. $\mu^\dagger$-distributed samples. In fact, there is a prior that gives positive mass to every open subset of Θ but yields consistent posterior estimates for only a first-category set of possible "true" (data-generating) parameter values $\theta^\dagger$.

There are conditions on priors that do ensure consistency in infinite-dimensional or non-parametric contexts, e.g. the tail-free priors introduced by Freedman in [54] and hybrid Bayesian–frequentist tools such as Dirichlet process priors [60]. However, while the collection of "bad" priors that lead to inconsistent results is measure-theoretically small [44, 33], it is topologically generic [55].

It is important to appreciate that the requirement of positive prior mass in every neighborhood of the true distribution depends upon the topology placed upon $\mathcal{M}(\mathcal{X})$. For example, Schwartz [96] showed that every π that puts positive mass on all Kullback–Leibler (relative entropy) neighborhoods of $\mu^\dagger$ is weakly consistent. On the other hand, Freedman [54] and Diaconis & Freedman [42] show that π may put positive mass on all weak neighborhoods of $\mu^\dagger$ and still fail to be weakly consistent — e.g. by not being tail-free. Nor are results limited to *weak* convergence of the posterior to $\mu^\dagger$. For example, [10] shows that consistency holds in the Hellinger distance if π puts positive mass on all Kullback–Leibler neighborhoods of $\mu^\dagger$ and certain smoothness and tail conditions are satisfied; see [112, 115] for further results on Hellinger and Kullback–Leibler consistency. The amount of prior probability mass that lies Kullback–Leibler-close to the truth, quantified using a notion called *thickness*, can be used to quantify the convergence properties of Bayes estimates [1, 2, 79]. However, in the infinite-dimensional contexts that are increasingly subject to Bayesian analyses, it is important to note that probability measures are "usually" mutually singular and "rarely" mutually absolutely continuous, and so the Kullback–Leibler neighbourhoods of $\mu^\dagger$ are small sets that are "unlikely" to intersect the model class.

The situation in which there is no $\theta^\dagger \in \Theta$ such that $\mu^\dagger = \mu(\cdot\,, \theta^\dagger)$ is referred to as *model misspecification*. The consistency and other asymptotic properties of misspecified models appear to have first been considered by Berk [21, 22] and Huber [67]. See [73, 74] for a recent contribution, and [79] for convergence rates.

> "In practice, Bayesian inference is employed under misspecification *all the time*, particularly so in machine learning applications. While sometimes it works quite well under misspecification [26, 73], there are also cases where it does not [35, 58], so it seems important to determine precise conditions under which misspecification is harmful — even if such an analysis is based on frequentist assumptions."

There is a reasonable popular belief that gross misspecification of the model will be detected by some means before engaging in a serious Bayesian analysis; indeed there do exist tests [63, 119] for model misspecification, but it is important to note that while one *can* determine that the model is misspecified, one *cannot* be sure that the model is well-specified. There is also an understandable popular belief that these tests mean that one need only be concerned with the situation of "mild misspecification", and that provided $\mu^\dagger$ lies "close enough" to the model class $\{\mu(\cdot\,, \theta)\}_{\theta \in \Theta}$ (as illustrated in schematic form in Figure 2.1), the posterior estimates will still converge to a usefully informative limit. Simply put, one aim of this paper is to show that this belief is wrong.

2.2 Bayesian Robustness

> "Most statisticians would acknowledge that an analysis is not complete unless the sensitivity of the conclusions to the assumptions is investigated. Yet, in practice, such sensitivity analyses are rarely used. This is because sensitivity analyses involve difficult computations that must often be tailored to the specific problem. This is especially true in Bayesian inference where the computations are already quite difficult." [116]

One response to the concern that the choice of prior (and likelihood) is somewhat arbitrary is to perform Bayesian analysis over classes of priors (and likelihoods): this approach is known as *robust Bayesian inference*. The robust Bayesian viewpoint appears to have been introduced independently by Box [29] and Huber [66]; see e.g. [19, 20] and Chapter 15 of [68] for surveys of the field. In the robust Bayesian approach, a class Π of priors and a class Λ of likelihoods together produce a class of posteriors by pairwise combination through Bayes' rule. Robust Bayesian methods are a subclass of the methods of *imprecise probability*; the idea that the probability of an event need not be a single real number has a history stretching back to Boole [28] and Keynes [72], with more recent and comprehensive foundations laid out in e.g. [75, 114, 118].

One way of generating such a class Π of priors is via a belief function, as in [117] and Dempster–Shafer theory more generally. The belief function framework encompasses prior probabilities whose values are known only on some finite partition of the probability space, and not the whole σ-algebra; classes of ε-contaminated priors can also be

represented in this way, as well as classes of locally perturbed priors. The belief function approach has the useful feature that explicit formulae can be given for the lower and upper posterior probabilities of events [117, Theorem 4.1].

Another typical approach to generating a class Π might be to consider a finite-dimensional parametrized class of models. For example, one could consider, instead of a single Gaussian prior on $\mathbb{R}$ of specified mean and variance, a two-parameter class of Gaussian priors with a range of means and variances, or a three-parameter class of skew-Gaussian priors. Similarly, one might consider a two-parameter class of beta distributions instead of a uniform prior on a bounded interval.

However, a danger in specifying a finite-dimensional class Π of priors is that one is making very strong statements about the form of the priors, particularly with regard to the tails, that cannot be justified based on often-limited amounts of prior information. For example, if all the priors $\pi \in \Pi$ have thin tails, then the class Π will have a very difficult time modelling events that lie in those tails, even when exposed to data from those regions. This problem is particularly important in applied fields such as catastrophe modelling, insurance, and re-insurance, in which the catastrophic events of interest are by definition high-impact low-probability "Black Swan" events: the difference between an exponentially small and an inverse-polynomially small tail can be vitally important. Also, because members of a finite-dimensional parametric family Π of priors often have similar qualitative properties (such as being mutually absolutely continuous), the apparently broader perspective does not not add much to the asymptotic posterior picture in terms of robust consistency, although it does provide a broader understanding given finitely many samples.

Rather than specifying a finite-dimensional Π, it is epistemologically more reasonable to specify a finite-*codimensional* Π, for example by specifying interval bounds on the expected values of finitely many observed test functions (i.e. generalized moment inequalities); this setting encompasses the finite-partition belief function framework mentioned above. Calculation of optimal prior and posterior bounds on quantities of interest is often an exercise in numerical optimization [25, 86, 98] rather than closed-form formulae.

2.3 Purpose

In terms of the above discussion, one purpose of this paper is to explore the extent to which one can simultaneously have *robust* Bayesian analyses that produce *consistent* answers, given that the models used (both priors and likelihoods) are certain to be *misspecified* to some degree. Can one be "just a little bit wrong" in terms of model misspecification? Our results suggest that the answer is strongly negative when "closeness" is measured in total variation and Prokhorov metrics: either one's robust Bayesian model is well-specified, in which case there is robust consistency; or else the model is misspecified — even slightly — and the limiting posterior bounds are no more informative than "L^∞-type" worst- and best-case bounds.

3 Incorporation of Bayesian Priors in the OUQ framework

3.1 A quick reminder on the OUQ framework

Let $\mathcal{X}$ be a topological space, let $\mathcal{F}(\mathcal{X})$ be the space of real-valued measurable functions on $\mathcal{X}$, and let $\mathcal{G} \subseteq \mathcal{F}(\mathcal{X})$ be a subset. Let $\mathcal{A}$ be an arbitrary subset of $\mathcal{G} \times \mathcal{M}(\mathcal{X})$, and let $\Phi\colon \mathcal{G} \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a function producing a quantity of interest. In the context of uncertainty quantification one is interested in estimating $\Phi(f^\dagger, \mu^\dagger)$, where $(f^\dagger, \mu^\dagger) \in \mathcal{G} \times \mathcal{M}(\mathcal{X})$ corresponds to an *unknown reality*: the function $f^\dagger$ represents a *response function* of interest, and $\mu^\dagger$ represents the probability distribution of the inputs of $f^\dagger$. If $\mathcal{A}$ represents all that is known about $(f^\dagger, \mu^\dagger)$ (in the sense that $(f^\dagger, \mu^\dagger) \in \mathcal{A}$ and that any $(f, \mu) \in \mathcal{A}$ could, a priori, be $(f^\dagger, \mu^\dagger)$ given the available information) then [86] shows that the quantities

$$\mathcal{U}(\mathcal{A}) := \sup_{(f,\mu)\in\mathcal{A}} \Phi(f, \mu) \tag{3.1}$$

$$\mathcal{L}(\mathcal{A}) := \inf_{(f,\mu)\in\mathcal{A}} \Phi(f, \mu) \tag{3.2}$$

determine the inequality

$$\mathcal{L}(\mathcal{A}) \leq \Phi(f^\dagger, \mu^\dagger) \leq \mathcal{U}(\mathcal{A}), \tag{3.3}$$

to be optimal given the available information $(f^\dagger, \mu^\dagger) \in \mathcal{A}$ as follows: It is simple to see that the inequality (3.3) follows from the assumption that $(f^\dagger, \mu^\dagger) \in \mathcal{A}$. Moreover, for any $\varepsilon > 0$ there exists a $(f, \mu) \in \mathcal{A}$ such that

$$\mathcal{U}(\mathcal{A}) - \varepsilon < \Phi(f, \mu) \leq \mathcal{U}(\mathcal{A}).$$

Consequently since all that we know about $(f^\dagger, \mu^\dagger)$ is that $(f^\dagger, \mu^\dagger) \in \mathcal{A}$, it follows that the upper bound $\Phi(f^\dagger, \mu^\dagger) \leq \mathcal{U}(\mathcal{A})$ is the best obtainable given that information, and the lower bound is optimal in the same sense.

Although the OUQ optimization problems (3.1) and (3.2) are extremely large, we have shown in [86] that an important subclass enjoys significant and practical finite-dimensional reduction properties. First, by [86, Cor. 4.4], although the optimization variables (f, μ) lie in a product space of functions and probability measures, for OUQ problems governed by linear inequality constraints on generalized moments, the search can be reduced to one over probability measures that are products of finite convex combinations of Dirac masses with explicit upper bounds on the number of Dirac masses.

Furthermore, in the special case that all constraints are generalized moments of functions of f, the dependency on the coordinate positions of the Dirac masses is eliminated by observing that the search over admissible functions reduces to a search over functions on an m-fold product of finite discrete spaces, and the search over m-fold products of finite convex combinations of Dirac masses reduces to a search over the products of probability measures on this m-fold product of finite discrete spaces [86, Thm. 4.7]. Finally, by [86, Thm. 4.9], using the lattice structure of the space of functions, the search over these functions can be reduced to a search over a finite set.

Example 3.1. A classic example is $\Phi(f, \mu) := \mu[f \geq a]$ where a is a safety margin. In the certification context one is interested in showing that $\mu^\dagger[f^\dagger \geq a] \leq \varepsilon$, where ε is a safety certification threshold (i.e. the maximum acceptable $\mu^\dagger$-probability of the system $f^\dagger$ exceeding the safety margin a). If $\mathcal{U}(\mathcal{A}) \leq \varepsilon$, then the system associated with $(f^\dagger, \mu^\dagger)$ is safe even in the worst case scenario (given the information represented by $\mathcal{A}$). If $\mathcal{L}(\mathcal{A}) > \varepsilon$, then the system associated with $(f^\dagger, \mu^\dagger)$ is unsafe even in the best case scenario (given the information represented by $\mathcal{A}$). If $\mathcal{L}(\mathcal{A}) \leq \varepsilon < \mathcal{U}(\mathcal{A})$, then the safety of the system cannot be decided (although we could declare the system to be unsafe due to lack of information).

3.2 Bayesian priors on spaces of measures and functions

In the OUQ setting, an assumption of the form

$$(f^\dagger, \mu^\dagger) \in \mathcal{A},$$

in terms of an assumption set $\mathcal{A} \subseteq \mathcal{G} \times \mathcal{M}(\mathcal{X})$ where $\mathcal{G} \subseteq \mathcal{F}(\mathcal{X})$, was used to derive the optimal inequality (3.3). This paper will consider the situation in which one has priors on the admissible set $\mathcal{A}$ and also information in the form of sample data. One of our goals is to analyse the robustness (or brittleness) of Bayesian inference by obtaining optimal bounds on posterior values given local misspecifications.

In order to define priors on the space of admissible scenarios, $\mathcal{A}$ needs to be given the structure of a measurable space; i.e. a suitable σ-algebra $\Sigma_{\mathcal{A}}$ on $\mathcal{A}$ must be provided. When this is accomplished we will refer to a probability measure $\pi \in \mathcal{M}(\Sigma_{\mathcal{A}})$ as a *prior*. However, this is a non-trivial task because if the σ-algebra on $\mathcal{A}$ is too small then natural functions of interest Φ may not be measurable, and if it is too large then the set of probability measures on $\mathcal{A}$ may be empty or too small. Moreover, it would also be convenient if we could easily apply the reduction theorems of [86].

Section 7 will show that Polish (i.e. separable and completely metrizable) spaces provide a natural setting for our work. In particular, we develop simple conditions on the function space $\mathcal{G}$ and base space $\mathcal{X}$ for which: (1) the reduction theorems of [86] apply when $\mathcal{A}$ is any Borel subset of the product space $\mathcal{G} \times \mathcal{M}(\mathcal{X})$, when $\mathcal{M}(\mathcal{X})$ is endowed with the weak-$*$ topology; and (2) the classic object of interest $\Phi\colon \mathcal{G} \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ defined by $\Phi(f, \mu) := \mu[f \geq a]$ is measurable. As stated in Theorem 7.5 the conditions are that $\mathcal{X}$ be Polish, $\mathcal{G}$ be Polish, and the evaluation function $i_x\colon \mathcal{G} \to \mathbb{R}$ defined by $i_x(g) := g(x)$ be Borel measurable for each $x \in \mathcal{X}$. Moreover, we show that many function spaces satisfy these conditions: Theorem 7.10 shows that a Reproducing Kernel Hilbert Space (RKHS) or Reproducing Kernel Banach Space (RKBS) of functions over $\mathcal{X}$ with measurable feature map(s) satisfy these criteria; Theorem 7.12 asserts that the space of upper semicontinuous functions with the Wijsman topology also satisfies these conditions. In addition, many of these function spaces are known to be very expressive. For example, Steinwart [100] introduced *universal kernels* on compact domains as those whose RKHSs which can approximate any continuous function uniformly, and demonstrated that many of the existing popular kernels, in particular the Gaussian kernels,

are universal. For non-compact $\mathcal{X}$, Steinwart, Hush and Scovel [103] provide conditions on the kernel that guarantee approximation properties in L^p spaces. For a thorough discussion of these matters in the context of Learning Theory, see [101].

In the process of establishing these results, we have obtained a new result which, in addition to being very useful to us, may be important for Learning Theory: on a bianalytic space (see Frolik [57] for the definition and the proof that a Polish space is bianalytic), Lemma 7.9 implies that a RKHS or RKBS with measurable primary feature map is separable.

Remark 3.2. The desire to have the Borel measurable structure of a Polish space might seem to be a spurious level of abstraction, but there are many good reasons for it. The first is that, by Suslin's Theorem [71, Thm. 14.2], all Borel subsets of a Polish space are Suslin, where a *Suslin space* is a continuous Hausdorff image of a Polish space. Indeed, Suslin sets are important in measurable selection theorems (see e.g. [34]) such as those that we use in the proof of Lemma 4.10; furthermore, in addition to Ulam's theorem [7, Thm. 4.3.8] that all probability measures on a Polish space are regular (approximable from within by compact sets), Schwartz' theorem [97] implies that that all probability measures on a Suslin space are regular, and, therefore, [108, Thm. 11.1] implies that the extreme points in the space of probability measures on a Suslin space are the Dirac measures. Consequently, when the product $\mathcal{G} \times \mathcal{M}(\mathcal{X})$ is Polish, any Borel subset $\mathcal{A} \subseteq \mathcal{G} \times \mathcal{M}(\mathcal{X})$ is Suslin and so the extreme points of probability measures on $\mathcal{A}$ are the Dirac measures, and some powerful measurable selection theorems are available. Moreover, when the base space is metrizable, then the space of probability measures is Polish in the weak-$*$ topology if and only if the base space is Polish.

Furthermore, since separability is equivalent to second countablility for metric spaces, we have that the Borel structure of a product is the product of Borel structures of Polish spaces. In addition, by [48, Thm. 10.2.2], regular conditional probabilities exist for observables with values in a Polish space. Moreover, in some sense, there is only one Polish measurable space by a construction of Skorokhod [69]. Also, Polish spaces are the spaces of Descriptive Set Theory, see e.g. Kechris [71], and fundamental to our results, see Lemma 7.2, will be a surprising result of Kechris. Finally, Polish spaces appear to be the appropriate spaces to play topological games such as the Banach–Mazur game [87], the Sierpiński game, the Ulam game, the Banach game, and the Choquet game. Moreover, Choquet's theorem [71, Thm. 8.18] says that a separable metric space is completely metrizable (and hence Polish) if and only if the second player has a winning strategy in the strong Choquet game. For a review of topological games, see Telgársky's review [107], and for topological games in hyperspace see that of Zsilinszky [124].

3.3 Data Spaces and Maps

Henceforth $\mathcal{A}$ will be a topological space. In practice the response function $f^\dagger$ and the probability measure $\mu^\dagger$ are not directly observed and the sample data arrives in the form of (realizations of) observation random variables, the distribution of which is related to $(f^\dagger, \mu^\dagger)$. To simplify the current presentation we will assume that this

relation is determined by a function of $(f^\dagger, \mu^\dagger)$ — such as the case where the data $(X_1, f^\dagger(X_1)), \ldots, (X_n, f^\dagger(X_n))$ are determined by n independent realizations X_i of the random variable X determined by the possibly unknown distribution $\mu^\dagger$. Throughout this paper we will use the following notation: $\mathcal{D}$ will denote the observable space (i.e. the space in which the sample data take values); $\mathcal{D}$ will be assumed to be a metrizable Suslin space and D will denote a $\mathcal{D}$-valued random variable producing the observed sample data. To represent the dependence of the observation random variable D on the unknown state $(f^\dagger, \mu^\dagger) \in \mathcal{A}$ we introduce a measurable function

$$\mathbb{D}\colon \mathcal{A} \to \mathcal{M}(\mathcal{D}),$$

where $\mathcal{M}(\mathcal{D})$ is given the Borel structure corresponding to the weak-$*$ topology, to define this relation. The idea is that $\mathbb{D}(f, \mu)$ is the probability distribution of the observed sample data $D(f, \mu)$ if $(f^\dagger, \mu^\dagger) = (f, \mu)$, and for this reason it may be called the *data map* or — even more loosely — the *observation operator*. Often, for simplicity, we will write D instead of $D(f, \mu)$. For simplicity and clarity, we save for Section 3.5 the consideration of the case where the sample process $\mathbb{D}(f, \mu)$ has uncertainties with (f, μ) known.

We proceed with a natural generalization of the Campbell measure and Palm distribution associated with a random measure as described in [70] (see also [38, Ch. 13] for a more current treatment). To that end, observe that since $\mathcal{D}$ is metrizable, it follows from [4, Thm. 15.13], that, for any $B \in \mathcal{B}(\mathcal{D})$, the evaluation $\nu \mapsto \nu(B)$, $\nu \in \mathcal{M}(\mathcal{D})$, is measurable. Consequently, the measurability of $\mathbb{D}$ implies that the mapping

$$\widehat{\mathbb{D}}\colon \mathcal{A} \times \mathcal{B}(\mathcal{D}) \to R$$

defined by

$$\widehat{\mathbb{D}}\big((f, \mu), B\big) := \mathbb{D}(f, \mu)[B], \quad \text{for } (f, \mu) \in \mathcal{A}, B \in \mathcal{B}(\mathcal{D})$$

is a transition function in the sense that, for fixed $(f, \mu) \in \mathcal{A}$, $\widehat{\mathbb{D}}\big((f, \mu), \cdot\,\big)$ is a probability measure, and, for fixed $B \in \mathcal{B}(\mathcal{D})$, $\widehat{\mathbb{D}}\big(\cdot\,, B\big)$ is Borel measurable. Therefore, by [27, Thm. 10.7.2], any $\pi \in \mathcal{M}(\mathcal{A})$, defines a probability measure

$$\pi \odot \mathbb{D} \in \mathcal{M}\big(\mathcal{B}(\mathcal{A}) \times \mathcal{B}(\mathcal{D})\big)$$

through

$$\pi \odot \mathbb{D}\big[A \times B\big] := \mathbb{E}_{(f.\mu)\sim\pi}\big[\mathbb{1}_A(f, \mu)\mathbb{D}(f, \mu)[B]\big], \quad \text{for } A \in \mathcal{B}(\mathcal{A}), B \in \mathcal{B}(\mathcal{D}), \tag{3.4}$$

where $\mathbb{1}_A$ is the indicator function of the set A:

$$\mathbb{1}_A(f, \mu) := \begin{cases} 1, & \text{if } (f, \mu) \in A, \\ 0, & \text{if } (f, \mu) \notin A. \end{cases}$$

It is easy to see that π is the $\mathcal{A}$-marginal of $\pi \odot \mathbb{D}$. Moreover, when $\mathcal{X}$ is Polish, [4, Thm. 15.15] implies that $\mathcal{M}(\mathcal{X})$ is Polish, and when $\mathcal{G}$ is Polish it follows that

$\mathcal{A} \subseteq \mathcal{G} \times \mathcal{M}(\mathcal{X})$ is second countable. Consequently, since $\mathcal{D}$ is Suslin and hence second countable, it follows from [48, Prop. 4.1.7] that

$$\mathcal{B}\big(\mathcal{A} \times \mathcal{D}\big) = \mathcal{B}(\mathcal{A}) \times \mathcal{B}(\mathcal{D})$$

and hence $\pi \odot \mathbb{D}$ is a probability measure on $\mathcal{A} \times \mathcal{D}$. That is,

$$\pi \odot \mathbb{D} \in \mathcal{M}(\mathcal{A} \times \mathcal{D}).$$

Let us refer to an element of $\mathcal{M}(\mathcal{A})$ as a *prior* on $\mathcal{A}$. With a prior π on $\mathcal{A}$, the quantity of interest $\Phi(f, \mu)$ becomes a random variable and we will be interested in estimating its distribution conditioned on the observation $D \in B$, where $B \in \mathcal{B}(\mathcal{D})$.

Example 3.3. In the context of Example 3.1, we are interested in estimating the probability (under the prior π) that the system is unsafe, conditioned on the observations $D \in B$, i.e. the conditional expectation

$$(\pi \odot \mathbb{D})\Big[\mu[f \geq a] > \epsilon \Big| D \in B\Big]$$

If D corresponds to observing independent realizations of $(X, G(X))$, then the observation space $\mathcal{D}$ is $(\mathcal{X} \times \mathbb{R})^n$ and the measure $\mathbb{D}(f, \mu)$ is the one associated with the random variable $D = \big((X^1, f(X^1)), \ldots, (X^n, f(X^n))\big)$ where the X^i are independent and distributed according to μ.

If D is the random variable that results from observing n independent realizations of $(X, f(X) + \xi)$ (f is observed with additive Gaussian noise $\xi \sim \mathcal{N}(0, \sigma^2)$), then the measure $\mathbb{D}(f, \mu)$ is the one associated with the random variable $D = \big((X^1, f(X^1) + \xi^1), \ldots, (X^n, f(X^n) + \xi^n)\big)$ where the X^i are independent and distributed according to μ and the ξ^i are independent Gaussian random variables of mean zero and variance σ^2.

3.4 Bayes' Theorem and conditional expectation

Henceforth $\mathcal{A}$ will be a Suslin space, and suppose now that we have $\pi \odot \mathbb{D} \in \mathcal{M}(\mathcal{A} \times \mathcal{D})$ constructed in the above way. Let $\pi \cdot \mathbb{D}$ denote the corresponding Bayes' sampling distribution defined by the $\mathcal{D}$-marginal of $\pi \odot \mathbb{D}$, and note that, by (3.4), we have

$$\pi \cdot \mathbb{D}[B] := \mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f, \mu)[B]\big], \quad \text{for } B \in \mathcal{B}(\mathcal{D}). \tag{3.5}$$

Since both $\mathcal{D}$ and $\mathcal{A}$ are Suslin it follows that the product $\mathcal{A} \times \mathcal{D}$ is Suslin. Consequently, [27, Cor. 10.4.6] asserts that regular conditional probabilities exist for any sub-σ-algebra of $\mathcal{B}\big(\mathcal{A} \times \mathcal{D}\big)$. In particular, the product theorem of [27, Thm. 10.4.11] asserts that product regular conditional probabilities

$$\big(\pi \odot \mathbb{D}\big)|_d \in \mathcal{M}(\mathcal{A}), \quad \text{for } d \in \mathcal{D}$$

exist and that they are $\pi \cdot \mathbb{D}$-a.e. unique.

When we consider $\pi \in \mathcal{M}(\mathcal{A})$ a prior, then this result can be interpreted as the posteriors of Bayes' Theorem. However, because such regular conditional probabilities

are only uniquely defined $\pi \cdot \mathbb{D}$-a.e., when a data sample $d \in \mathcal{D}$ arrives such that $\pi \cdot \mathbb{D}[\{d\}] = 0$, a posterior $(\pi \odot \mathbb{D})|_d$ that could be *any* of the $\pi \cdot \mathbb{D}$-a.e.-equal regular conditional probabilities evaluated at d appears to have dubious utility. Indeed, the fact that the regular conditional probabilities are only uniquely defined $\pi \cdot \mathbb{D}$-a.e. suggests that integrals of posteriors over subsets $B \in \mathcal{B}(\mathcal{D})$ such that $\pi \cdot \mathbb{D}[B] > 0$ are the more natural objects. Moreover, the restriction that B be an open set is natural for practical reasons, since conditioning on D lying in an open subset B rather than on its exact value is what one has to do when the sample data can only be observed after rounding error. Furthermore, we will show in Section 5 that if the data d have been sampled from a probability measure $\pi^\dagger \cdot \mathbb{D}$ for some $\pi^\dagger \in \mathcal{M}(\mathcal{A})$ (commonly called a "true prior" in Bayesian statistics) then with $\pi^\dagger \cdot \mathbb{D}$ probability one (on the realization of d), the $\pi^\dagger \cdot \mathbb{D}$-measure of any open set containing d is strictly positive. In other words, $\pi^\dagger \cdot \mathbb{D}$-almost surely, $\pi^\dagger$ (the "true prior") belongs to the random subset of $\mathcal{M}(\mathcal{A})$ defined as the priors $\pi \in \mathcal{M}(\mathcal{A})$ such that $\pi \cdot \mathbb{D}[B] > 0$ for any open set B containing the data d (this subset is randomized through the realization of the data d).

Finally, throughout, we will find it useful to assume that

$$\Phi \text{ is semibounded} \tag{3.6}$$

in that it is either bounded above or bounded below. Semiboundedness is sufficient to ensure that the integral of Φ with respect to any probability measure exists, possibly with the value ∞ or $-\infty$, and such integrands are sufficient for the reduction theorems of Winkler [121] that we use.

Remark 3.4. Note that the assumption that Φ is semibounded is mostly for convenience since integrands which are not semibounded, like that defining the first moment, can be considered by restricting the space of measures to those measures that have well defined first moments.

3.5 Incompletely specified priors and observation maps

In practical situations, the observation map $\mathbb{D}$ may be imperfectly known. For example:

1. the sample data may be corrupted with experimental noise with unknown distribution;
2. the observations of $(X, f^\dagger(X))$ may not be independent and the available information on their correlations may be limited to that contained in a covariance matrix;
3. the sample data may also not correspond to direct observations of $(X, f^\dagger(X))$ under the measure $\mu^\dagger$ but to an observation of random variables correlated through a unknown process, possibly involving an inverse problem (that may be ill-posed).

Let us also observe that: (1) the choice of a particular prior on $\mathcal{A}$ involves a degree of arbitrariness that may be incompatible with the certification of rare/critical events (2) the definition of such a prior is a non trivial task if $\mathcal{A}$ is infinite dimensional. For these reasons it is necessary to consider situations in which the prior π and the observation map $\mathbb{D}$ are imperfectly known or specified. More precisely, the (lack of) information (or specification) on π and $\mathbb{D}$ can be represented via the introduction of two spaces Π and $\mathfrak{D}$

where the subset $\Pi \subseteq \mathcal{M}(\mathcal{A})$ consists of the set of admissible priors π and $\mathfrak{D}$ is a subset of the set of all measurable mappings $\mathbb{D}$ from $\mathcal{A}$ into $\mathcal{M}(\mathcal{D})$.

One of our goals in allowing incompletely specified priors is to assess the robustness of posterior Bayesian estimates with respect to the particular choice of priors. More precisely we will compute optimal bounds on $\mathbb{E}_\pi[\Phi]$ when $\pi \in \Pi$ and show how these bounds are affected by the introduction of sample data by computing optimal bounds on $\mathbb{E}_{\pi \odot \mathbb{D}}[\Phi|B]$, for $B \in \mathcal{B}(\mathcal{D})$.

4 Optimal bounds on the prior value

Recall that for a subset $\mathcal{A}$ and a measurable quantity of interest $\Phi\colon \mathcal{A} \to \mathbb{R}$, that under the assumption $(f^\dagger, \mu^\dagger) \in \mathcal{A}$, we have the optimal upper $\mathcal{U}(\mathcal{A})$ and lower $\mathcal{L}(\mathcal{A})$ bounds on the *value* $\Phi(f^\dagger, \mu^\dagger)$ of the quantity of interest, defined in (3.1) and (3.2) by

$$\begin{aligned} \mathcal{U}(\mathcal{A}) &:= \sup_{(f,\mu)\in\mathcal{A}} \Phi(f,\mu) \\ \mathcal{L}(\mathcal{A}) &:= \inf_{(f,\mu)\in\mathcal{A}} \Phi(f,\mu)\,. \end{aligned}$$

When we put a prior π on $\mathcal{A}$, we have to define the *value* $\bar{\Phi}(\pi)$ *of the prior* π corresponding to an extended quantity $\bar{\Phi}\colon \mathcal{M}(\mathcal{A}) \to \mathbb{R}$ of interest corresponding to Φ. Disregarding integrability concerns, for a given Φ, let us call the induced function

$$\bar{\Phi}(\pi) := \mathbb{E}_\pi[\Phi], \quad \pi \in \mathcal{M}(\mathcal{A}), \tag{4.1}$$

the canonical one associated with Φ and abuse notation by denoting the function $\bar{\Phi}$ as Φ. For such a canonical quantity of interest, we call the value $\mathbb{E}_\pi[\Phi]$ the *prior value*, and note that the values

$$\mathcal{U}(\Pi) := \sup_{\pi\in\Pi} \mathbb{E}_\pi\big[\Phi\big] \tag{4.2}$$

$$\mathcal{L}(\Pi) := \inf_{\pi\in\Pi} \mathbb{E}_\pi\big[\Phi\big] \tag{4.3}$$

form a natural generalization of the values $\mathcal{U}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A})$. Moreover, in the same way that $\mathcal{U}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A})$ are optimal upper and lower bounds on $\Phi(f^\dagger, \mu^\dagger)$ given the information that $(f^\dagger, \mu^\dagger) \in \mathcal{A}$, $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$ are optimal upper and lower bounds on $\mathbb{E}_\pi\big[\Phi\big]$ given the information that $\pi \in \Pi$. Of course, to be well defined, integrability concerns should be addressed. Indeed, Assumption 3.6 that Φ is semibounded implies that $\mathbb{E}_\pi\big[\Phi\big]$ is well defined for any bounded measure π, possibly with the value ∞ and $-\infty$, and therefore the quantities in (4.2) and (4.3) are well defined.

Remark 4.1. The restriction that the the extended quantity of interest corresponding to Φ be canonical is really no restriction, but is assumed only to simplify the presentation and notation. Indeed, there are many important extended quantities of interest that are not affine as functions of the measure π. However, all the ones that we have

thought of can be handled by small modifications of the present framework, and their inclusion here would simply complicate the presentation and notation. Moreover, note that many affine non-canonical extended quantities of interest become canonical through simple transformations. For example, when $\Phi_1(f,\mu) := \mu[f \geq a]$ is a quantity of interest, and the extended quantity of interest is the probability that the system is unsafe, i.e. $\pi(\{(f,\mu) \mid \mu[f \geq a] > \varepsilon\})$ where $\{(f,\mu) \mid \mu[f \geq a] > \varepsilon\}$ is the set of unsafe (f,μ), then this extended quantity of interest is not canonical with respect to Φ_1. However, by transformation to $\Phi_2 := \mathbb{1}_{\{r|r>\varepsilon\}} \circ \Phi_1$, the extended quantity of interest becomes canonical and $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$, defined in terms of Φ_2, are optimal upper and lower bounds on the probability that the system is unsafe given the set of priors Π.

4.1 General information barriers on prior values

Let $\delta\colon \mathcal{A} \to \mathcal{M}(\mathcal{A})$ be the mapping of points to unit Dirac measures, where we denote $\delta_{(f,\mu)}$ as the Dirac mass at (f,μ), and, for $\Pi \subseteq \mathcal{M}(\mathcal{A})$, define

$$\mathcal{A}_\Pi := \delta^{-1}\Pi = \{(f,\mu) \in \mathcal{A} \mid \delta_{(f,\mu)} \in \Pi\}. \tag{4.4}$$

That is, $\mathcal{A}_\Pi$ consists of those scenarios (f,μ) that are not only admissible in the sense that they lie in $\mathcal{A}$, but are also admissible as a prior in the sense that $\delta_{(f,\mu)}$ is an element of Π.

With the convention that $\mathcal{U}(\varnothing) := -\infty$ and $\mathcal{L}(\varnothing) := +\infty$, the following theorem shows the relationships among $\mathcal{U}(\mathcal{A})$ and $\mathcal{U}(\mathcal{A}_\Pi)$ as defined by (3.1), $\mathcal{L}(\mathcal{A})$ and $\mathcal{L}(\mathcal{A}_\Pi)$ as defined by (3.2), and $\mathcal{U}(\Pi)$ and $\mathcal{L}(\Pi)$ as defined by (4.2) and (4.3).

Theorem 4.2. *It holds true that*

$$\mathcal{U}(\mathcal{A}_\Pi) \leq \mathcal{U}(\Pi) \leq \mathcal{U}(\mathcal{A})$$

and

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi) \leq \mathcal{L}(\mathcal{A}_\Pi).$$

Moreover, if $\mathcal{A}_\Pi$ is non-empty, then

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi) \leq \mathcal{L}(\mathcal{A}_\Pi) \leq \mathcal{U}(\mathcal{A}_\Pi) \leq \mathcal{U}(\Pi) \leq \mathcal{U}(\mathcal{A}).$$

4.2 Priors specified through marginals

In many settings, probability measures or sets of probability measures are specified through generalized moments or other properties of marginal distributions. To analyse this case, let $\mathcal{Q}$ be a topological space and consider a measurable map $\Psi\colon \mathcal{A} \to \mathcal{Q}$. Let us abuse notation by also denoting the corresponding pushforward of measures $\Psi : \mathcal{M}(\mathcal{A}) \to \mathcal{M}(\mathcal{Q})$ by the same symbole Ψ. For a probability measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$, let

$$\Psi^{-1}\mathbb{Q} := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi = \mathbb{Q}\}$$

be the set of probability measures $\pi \in \mathcal{M}(\mathcal{A})$ that push forward to $\mathbb{Q}$. More generally, for a non-empty set $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$, let

$$\Psi^{-1}\mathfrak{Q} := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\} \tag{4.5}$$

be the set of probability measures $\pi \in \mathcal{M}(\mathcal{A})$ such that $\Psi\pi \in \mathfrak{Q}$. Now, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be an admissible set of Ψ-marginals. Then the corresponding admissible set of priors is $\Psi^{-1}\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{A})$ and the corresponding objects to be computed are $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ and $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ according to (4.2) and (4.3).

We will now demonstrate how to reduce the computation of $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ and $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ when $\mathfrak{Q}$ is specified by linear inequalities. Later, in Section 4.2.2, we will develop a more powerful *nested* reduction which will provide the foundation for our reduction methods.

Before we begin, we need to introduce some terminology. Following Winkler [121], let $\mathcal{Y}$ be a topological space and let $\mathcal{M} \subseteq \mathcal{M}(\mathcal{Y})$ be a set of measures. Let $\mathrm{ext}(\mathcal{M})$ denote the set of extreme points of $\mathcal{M}$ and let the evaluation field $\Sigma(\mathrm{ext}(\mathcal{M}))$ be the smallest σ-algebra of subsets of $\mathrm{ext}(\mathcal{M})$ such that the evaluation map $\nu \mapsto \nu(B)$ is measurable for all $B \in \mathcal{B}(\mathcal{Y})$. Then a measure $\nu \in \mathcal{M}(\mathcal{Y})$ is said to be a *barycenter* of $\mathcal{M}$ if there exists a probability measure p on $\Sigma(\mathrm{ext}(\mathcal{M}))$ such that the *barycentric formula*

$$\nu(B) = \int_{\mathrm{ext}(\mathcal{M})} \nu'(B)\,\mathrm{d}p(\nu'), \quad B \in \mathcal{B}(\mathcal{Y}) \tag{4.6}$$

holds. Furthermore, the following notion of a *measure affine function* is central to Winkler's [121] reduction theorems, which we use:

Definition 4.3. An extended real-valued function F on $\mathcal{M} \subseteq \mathcal{M}(\mathcal{Y})$ is said to be *measure affine* if, for all $\nu \in \mathcal{M}$ and all probability measures p on $\Sigma(\mathrm{ext}(\mathcal{M}))$ for which the barycentric formula (4.6) holds, F is p-integrable and

$$F(\nu) = \int_{\mathrm{ext}(\mathcal{M})} F(\nu')\,\mathrm{d}p(\nu').$$

A major consequence of the assumption (3.6), that Φ is semibounded, is that $\mathbb{E}_\nu[\Phi]$ exists, with possible values ∞ and $-\infty$, for all finite measures ν. As a consequence, by [121, Prop. 3.1], the extended-real-valued function

$$\nu \mapsto \mathbb{E}_\nu[\Phi]$$

is measure affine.

4.2.1 Primary reduction for prior values

Let us consider the computation of

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_\pi[\Phi] \tag{4.7}$$

when $\mathfrak{Q}$ is specified by n generalized moment inequalities determined by measurable functions g_i $i = 1, \ldots, n$. The situation for the lower bound $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$ is the same. That is, let $I_i, i = 1, \ldots, n$ be n closed intervals, allowing semi-infinite intervals $(-\infty, q_i]$ and $[q_i, \infty)$, and define

$$\mathfrak{Q} = \{\mathbb{Q} \in \mathcal{M}(\mathcal{Q}) \,|\, \mathbb{E}_{\mathbb{Q}}[g_i] \in I_i \text{ for } i = 1, \ldots, n\},$$

where implicit in the definition is that all n integrals exist. Then, by a change of variables, $\mathbb{E}_{\Psi\pi}[g_i] = \mathbb{E}_{\pi}[g_i \circ \Psi]$ holds if either integral exists (see e.g. [11, Cor. 19.2]), so we conclude that

$$\begin{aligned}\Psi^{-1}\mathfrak{Q} &:= \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\} \\ &= \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\Psi\pi}[g_i] \in I_i \text{ for } i = 1, , n\} \\ &= \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\pi}[g_i \circ \Psi] \in I_i \text{ for } i = 1, \ldots, n\}\end{aligned}$$

and so conclude that $\Psi^{-1}\mathfrak{Q}$ is defined by the n generalized moment inequalities corresponding to $g_i \circ \Psi \colon \mathcal{A} \to \mathbb{R}$ for $= 1, \ldots, n$. Consequently, since the function $\pi \mapsto \mathbb{E}_{\pi}[\Phi]$ is measure affine, it follows from the reduction theorems of [86] that we can reduce the supremum on the right-hand side of (4.7) to the convex combination of $n + 1$ Dirac masses. To state the theorem we have just proven, let

$$\Delta(n) := \left\{ \sum_{i=0}^{n} \alpha_i \delta_{(f_i, \mu_i)} \;\middle|\; (f_i, \mu_i) \in \mathcal{A}, \alpha_i \geq 0, \text{for } i = 0, \ldots, n \right\}. \tag{4.8}$$

be the set of non-negative combinations of $n + 1$ Dirac masses. Let the vector I of intervals have components I_i for $i = 1, \ldots, n$, let

$$\Pi(I) := \Psi^{-1}\mathfrak{Q}$$

be defined as above, and consider the subset

$$\Pi(I, n) := \Pi(I) \cap \Delta(n) \subseteq \Pi(I) \tag{4.9}$$

of those measures which are the $n + 1$-fold convex combinations of Dirac masses.

Theorem 4.4. *Let $\mathcal{A}$ be Suslin, let $\mathcal{Q}$ be separable and metrizable, and let $\Psi \colon \mathcal{A} \to \mathcal{Q}$ be measurable. Moreover, for n measurable functions $g_1, \ldots, g_n \colon \mathcal{Q} \to \mathbb{R}$ and n closed intervals $I_1, \ldots, I_n$, let*

$$\mathfrak{Q} := \{\mathbb{Q} \in \mathcal{M}(\mathcal{Q}) \,|\, \mathbb{E}_{\mathbb{Q}}[g_i] \in I_i \text{ for } i = 1, \ldots, n\}$$

define the admissible set of Ψ-marginals. Then,

$$\mathcal{U}\big(\Pi(I)\big) = \mathcal{U}\big(\Pi(I, n)\big)$$

where

$$\mathcal{U}\big(\Pi(I, n)\big) = \begin{cases} \sup \sum_{i=0}^{n} \alpha_i \Phi(f_i, \mu_i) \\ \text{among } (f_i, \mu_i) \in \mathcal{A},\ \alpha_i \geq 0,\ \sum_{i=0}^{n} \alpha_i = 1 \\ \text{such that } \sum_{i=0}^{n} \alpha_i g_j\big(\Psi(f_i, \mu_i)\big) \in I_j \text{ for } j = 1, \ldots, n. \end{cases} \tag{4.10}$$

Remark 4.5. The freedom to determine intervals I_i, $i = 1, \dots, n$, is one way to incorporate uncertainty and maintain a reduction to $n+1$ Dirac masses. In particular, by choosing semi-infinite intervals $I_i := (-\infty, q_i]$ we obtain a reduction to $n+1$ Dirac masses for inequality constraints of the form $\mathbb{E}_{\mathbb{Q}}[g_i] \leq q_i$, and by choosing point intervals $I_i := [q_i, q_i]$ we obtain a reduction to $n+1$ Dirac masses for equality constraints of the form $\mathbb{E}_{\mathbb{Q}}[g_i] = q_i$. Moreover, by choosing the interval to be semi-infinite or point interval depending on the index i we obtain a reduction to $n+1$ Dirac masses for mixed equality and inequality constraints.

Theorem 4.4 can be put into a canonical form in the following way: by considering the modified feature map $\Psi' \colon \mathcal{A} \to \mathbb{R}^n$ with components

$$\Psi'_i := g_i \circ \Psi, \quad \text{for } i = 1, \dots, n,$$

it follows from the above that

$$\Psi^{-1}\mathfrak{Q} = \left\{ \pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_\pi[\Psi'] \in I \right\}.$$

That is, by changing from the feature map Ψ to Ψ' we end up with a constraint set defined by the first moment of the vector function Ψ'. Therefore, let us remove the $'$ from Ψ', and require $\Psi \colon \mathcal{A} \to \mathbb{R}^n$ to be measurable. The following theorem is the canonical form of Theorem 4.4. It is a corollary of Theorem 4.4 for the constraint $\mathbb{E}_\pi[\Psi] \in Z$ when $Z = I$ is a closed rectangle. However, it is true for arbitrary $Z \subseteq \mathbb{R}^n$.

Theorem 4.6. *Let $\mathcal{A}$ be Suslin, let $\Psi \colon \mathcal{A} \to \mathbb{R}^n$ be measurable, let $Z \subset \mathbb{R}^n$, and let*

$$\mathfrak{Q} := \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^n) \,|\, \mathbb{E}_{Q \sim \mathbb{Q}}[Q] \in Z\} \tag{4.11}$$

be the set of those measures whose first moment belongs to Z. Then, for

$$\Pi(Z) := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] \in Z\} \tag{4.12}$$

and $\Pi(Z, n) := \Pi(Z) \cap \Delta(n)$, we have

$$\mathcal{U}\big(\Pi(Z)\big) = \mathcal{U}\big(\Pi(Z, n)\big)$$

where

$$\mathcal{U}\big(\Pi(Z, n)\big) = \begin{cases} \sup \sum_{i=0}^n \alpha_i \Phi(f_i, \mu_i) \\ \textit{among } (f_i, \mu_i) \in \mathcal{A}, \alpha_i \geq 0, \sum_{i=0}^n \alpha_i = 1 \\ \textit{such that } \sum_{i=0}^n \alpha_i \Psi(f_i, \mu_i) \in Z. \end{cases} \tag{4.13}$$

Example 4.7. Let $\mathcal{X} := [0, 1]$, $\mathcal{Q} = \mathbb{R}$ and consider the admissible set $\mathcal{A} := \mathcal{M}([0, 1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$ for some $a \in (0, 1)$, and the map $\Psi \colon \mathcal{A} \to \mathbb{R}$ defined by $\Psi(\mu) := \mathbb{E}_\mu[X]$. Take as the set of admissible priors π on $\mathcal{A}$ the collection

$$\Pi := \left\{ \pi \in \mathcal{M}(\mathcal{A}) \,\middle|\, \mathbb{E}_{\mu \sim \pi}\big[\mathbb{E}_\mu[X]\big] = q \right\}$$

for some fixed $q \in (0, a)$. Then we will show that

$$\mathcal{U}(\Pi) = q/a. \tag{4.14}$$

To that end, observe that since $\mathbb{E}_{\mu\sim\pi}\big[\mathbb{E}_\mu[X]\big] = \mathbb{E}_\pi[\Psi]$, it follows that

$$\Pi = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] = q\}\,,$$

so that Theorem 4.6 implies that we can reduce the optimization in $\mathcal{U}(\Pi)$ to the supremum over $\mu_1, \mu_2 \in \mathcal{A}$, $\alpha \in [0, 1]$ of

$$\alpha\mu_1[X \geq a] + (1 - \alpha)\mu_2[X \geq a]$$

subject to the constraint

$$\alpha\mathbb{E}_{\mu_1}[X] + (1 - \alpha)\mathbb{E}_{\mu_2}[X] = q.$$

Introducing the slack variables $q_1 := \mathbb{E}_{\mu_1}[X]$, $q_2 := \mathbb{E}_{\mu_2}[X]$ and using [86, Thm. 4.1] to reduce this problem further in μ_1, μ_2, we obtain that $\mathcal{U}(\Pi)$ is equal to the supremum over $\alpha \in [0, 1]$ and $q_1, q_2 \in [0, 1]$ of

$$\alpha \min\{1, \tfrac{q_1}{a}\} + (1 - \alpha) \min\{1, \tfrac{q_2}{a}\}$$

subject to the constraint $\alpha q_1 + (1 - \alpha)q_2 = q$. Observing that the supremum is achieved at $q_1, q_2 \leq a$, we conclude that $\mathcal{U}(\Pi) = q/a$, establishing (4.14). Moreover, note that $\mathcal{U}(\Pi) = \mathcal{U}(\mathcal{A}_\Pi)$ for $\mathcal{A}_\Pi$ defined in (4.4) instead of the general inequality $\mathcal{U}(\mathcal{A}_\Pi) \leq \mathcal{U}(\Pi)$ of Theorem 4.2.

4.2.2 Nested reduction for prior values

The result of Example 4.7 can also be deduced through a *nested* reduction that we will find generally more useful for two reasons. The first is that, in practice, not only is it highly non-trivial to specify a prior on the space $\mathcal{A}$, since it requires quantifying information on an infinite-dimensional space, but it may also be undesirable to do so. Indeed, if an expert does not have a prior on the full space $\mathcal{A}$ but only on some projection $\Psi(\mathcal{A}) = \mathcal{Q}$, then, rather than arbitrarily picking one particular prior on the space $\mathcal{A}$ compatible with the specified prior on $\Psi(\mathcal{A})$, it might be preferable to work with the set of priors on $\mathcal{A}$ specified through such marginals. Our second and main motivation is that, even when we can do the reduction on the primary space $\mathcal{M}(\mathcal{A})$, the reduced space remains so large that it may not be amenable to computation. However with the nested reduction theorems given below, the reduced space becomes computationally manageable when $\mathcal{Q}$ is finite dimensional.

Example 4.8. A simple example is $\Phi(f, \mu) := \mu[f \geq a]$ (where a is a safety margin), $\Psi(f, \mu) = (\mathbb{E}_\mu[f], \mathrm{Var}_\mu[f])$, $\mathcal{Q} = \mathbb{R}^2$, $\mathfrak{Q} = \{\mathbb{Q}\}$ where $\mathbb{Q}$ corresponds to the uniform distribution on $[-1, 1] \times [3, 4]$. In that example, the expert has only "the prior" that the mean of f with respect to μ is uniformly distributed on $[-1, 1]$ and that the variance

of f with respect to μ is independent of its mean and uniformly distributed on $[3,4]$. Observe that in this situation $\mathfrak{Q}$ does not uniquely specify a prior $\pi \in \mathcal{M}(\mathcal{A})$ but an infinite-dimensional set of priors $\Psi^{-1}(\mathfrak{Q}) \subseteq \mathcal{M}(\mathcal{A})$ and a robust approach would require assessing the safety of the system under the whole set $\Psi^{-1}(\mathfrak{Q})$ rather than under a particular element π of that set.

Idea of the nested reduction. Roughly, the idea of the nested reduction is as follows. To compute (4.7), consider the induced function

$$\mathcal{U} \circ \Psi^{-1} \colon \mathcal{Q} \to \mathbb{R}$$

defined by

$$\big(\mathcal{U} \circ \Psi^{-1}\big)(q) := \mathcal{U}\big(\Psi^{-1}(q)\big) = \sup_{(f,\mu)\in\Psi^{-1}(q)} \Phi(f,\mu), \quad \text{for } q \in \mathcal{Q},$$

where we use the notation of (3.1). From this it is natural to consider

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}], \quad \text{for } \mathbb{Q} \in \mathfrak{Q}.$$

Let $\mathbb{Q} \in \mathfrak{Q}$. Then, for any π such that $\Psi\pi = \mathbb{Q}$, it follows that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\ &= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi]\end{aligned}$$

Unfortunately, it is not true that $\mathcal{U} \circ \Psi^{-1} \circ \Psi = \Phi$; instead it is $\big(\mathcal{U} \circ \Psi^{-1} \circ \Psi\big)(f,\mu) = \sup_{(f',\mu'):\Psi(f',\mu')=\Psi(f,\mu)} \Phi(f',\mu')$. However, if it were true, then we would obtain

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\ &= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi] \\ &= \mathbb{E}_{\pi}[\Phi]\end{aligned}$$

and conclude that

$$\sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] = \sup_{\pi\in\Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] = \mathcal{U}(\Psi^{-1}\mathfrak{Q}).$$

We will show that, despite the fact that $\mathcal{U} \circ \Psi^{-1} \circ \Psi \neq \Phi$, the conclusion

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \tag{4.15}$$

is still valid, provided that it is interpreted correctly. Heuristically, the reason for this is that the supremum $\sup_{\pi\in\Psi^{-1}\mathfrak{Q}}$ in $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$ is exploring the maximum value of Φ on level sets of Ψ very much like the supremum in $\big(\mathcal{U} \circ \Psi^{-1}\big)(q) = \sup_{\Psi^{-1}(q)} \Phi$.

If $\mathcal{A}$ is such that a reduction theorem, e.g. from [86], applies to reduce the computation of the inner supremum in $\mathcal{U} \circ \Psi^{-1}$ to the supremum over convex combinations

of Dirac masses, and the admissible set $\mathfrak{Q}$ is such that a reduction theorem applies to the computation of the outer supremum of $\sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U}\circ\Psi^{-1}]$, then the identity (4.15) represents a nesting of reductions.

Let us now establish a result like (4.15). To do so will require addressing three questions: (1) What kind of function is $\mathcal{U}\circ\Psi^{-1}$? (2) What kind of measures $\mathbb{Q}\in\mathcal{M}(\mathcal{Q})$ can define an integral of a function with properties discovered from the answer to (1)? (3) Can we obtain a measurable solution operator to the optimization problem $\big(\mathcal{U}\circ\Psi^{-1}\big)(q)$, where $q\in\mathcal{Q}$? To that end, let us first recall a definition of universally measurable functions.

Definition 4.9. Let $(T,\mathcal{T})$ be a measurable space, and for a positive measure ν on $(T,\mathcal{T})$, let $\mathcal{T}_\nu$ denote the ν-completion of $\mathcal{T}$. Let $\hat{\mathcal{T}} := \bigcap_\nu \mathcal{T}_\nu$, where the intersection is over all positive bounded measures ν, denote the universally measurable sets. A $\hat{\mathcal{T}}$-measurable function is said to be *universally measurable*.

At the heart of the commutative representation used for the nested reduction is the following optimal measurable selection lemma answering questions (1) and (3) above:

Lemma 4.10. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon\mathcal{A}\to\mathcal{Q}$ be measurable. Then, for any subset $T\subseteq\Psi(\mathcal{A})$,*

1. *$\mathcal{U}\circ\Psi^{-1}$ is $\hat{\mathcal{B}}(T)$-measurable*
2. *for all $\delta>0$, there exists a δ-optimal $\hat{\mathcal{B}}(T)$-measurable section of Ψ; that is, a $\hat{\mathcal{B}}(T)$-measurable function $\psi\colon T\to\mathcal{A}$ such that $\Psi\big(\psi(q)\big)=q$ for all $q\in T$ and*

$$\Phi\big(\psi(q)\big) > \mathcal{U}\big(\Psi^{-1}(q)\big) - \delta, \quad \text{for all } q\in T.$$

To answer question (2) above, define a *support* $\operatorname{supp}(\mathbb{Q})$ of a measure $\mathbb{Q}\in\mathcal{M}(\mathcal{Q})$, as in [4, Ch. 12.3], to be a closed set such that

- $\mathbb{Q}\big(\mathcal{Q}\setminus\operatorname{supp}(\mathbb{Q})\big)=0$, and
- if $G\subseteq\mathcal{Q}$ is open and $G\cap\operatorname{supp}(\mathbb{Q})\neq\varnothing$, then $\mathbb{Q}(G\cap\operatorname{supp}(\mathbb{Q}))>0$.

When $\mathcal{Q}$ is a separable and metrizable space, it follows that it is second countable and therefore, by [4, Thm. 12.14], all $\mathcal{Q}\in\mathcal{M}(\mathcal{Q})$ have a uniquely defined support. Now consider a measure $\mathbb{Q}\in\mathcal{M}(\mathcal{Q})$ such that $\operatorname{supp}(\mathbb{Q})\subseteq\Psi(\mathcal{A})$. Then, by Lemma 4.10, $\mathcal{U}\circ\Psi^{-1}$ is $\hat{\mathcal{B}}(\operatorname{supp}\mathbb{Q})$-measurable. Therefore, the expected value $\mathbb{E}_{\mathbb{Q}}[\mathcal{U}\circ\Psi^{-1}]$ can be defined by integration with respect to the completion $\hat{\mathbb{Q}}$:

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U}\circ\Psi^{-1}] := \mathbb{E}_{\hat{\mathbb{Q}}}[\mathcal{U}\circ\Psi^{-1}]. \tag{4.16}$$

More generally, for any universally measurable function f and any finite measure $\mathbb{Q}$, we define the expected value $\mathbb{E}_{\mathbb{Q}}[f]$ of f by

$$\mathbb{E}_{\mathbb{Q}}[f] := \mathbb{E}_{\hat{\mathbb{Q}}}[f]. \tag{4.17}$$

Such a method of defining integrals of, possibly non-Borel measurable, but universally measurable, functions brings up many questions such as: when is it uniquely defined?; for a fixed integrand, when is the expectation operation affine in the measure?; does it have a change a variables formula? All such questions have nice answers and, although we are sure that this is classical, we cannot find a reference for these facts so we have included statements and proofs of the facts needed in this paper in Section 9.1 of the Appendix.

We now state our nested reduction theorem of the form (4.15):

Theorem 4.11. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ measurable. Moreover, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ for all $\mathbb{Q} \in \mathfrak{Q}$. Then, for each $\mathbb{Q} \in \mathfrak{Q}$, $\Psi^{-1}\mathbb{Q}$ is non-empty. Moreover, the upper bound $\mathcal{U}(\Psi^{-1}\mathfrak{Q})$, defined in (4.2), satisfies*

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}\big[\mathcal{U} \circ \Psi^{-1}\big]. \tag{4.18}$$

where the expectations on the right-hand side are defined as in (4.16). Finally, the expectation operator on the right-hand side is measure affine in $\mathbb{Q}$.

Remark 4.12. Since the right-hand side is measure affine in $\mathbb{Q}$, if $\mathbb{Q}$ is specified through (multi-)linear generalized moment inequalities, then the reduction theorems of [86] can be applied to obtain the supremum over $\mathbb{Q}$ by reducing $\mathbb{Q}$ to a convex combination of a finite number of Dirac masses on $\mathcal{Q}$. Moreover, if $\mathfrak{Q}$ consists of a single element, i.e. $\mathfrak{Q} = \{\mathbb{Q}\}$, then

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \mathcal{U}(\Psi^{-1}\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\big[\mathcal{U} \circ \Psi^{-1}\big], \tag{4.19}$$

and the right hand-side of (4.19) can be evaluated via Monte Carlo sampling of $q \in \mathcal{Q}$ according to the measure $\mathbb{Q}$.

Remark 4.13. A similar theorem can obtained for the optimal lower bound $\mathcal{L}(\Psi^{-1}\mathfrak{Q})$. Throughout this paper, results given for optimal upper bounds $\mathcal{U}$ can be translated into results for optimal lower bounds $\mathcal{L}$ by considering the negative quantity of interest $-\Phi$ and for the sake of concision we will not write those results unless necessary.

Example 4.14. Consider again Example 4.7, where $\mathcal{X} := [0,1]$, $\mathcal{Q} = \mathbb{R}$, the admissible set $\mathcal{A} := \mathcal{M}([0,1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$ for some $a \in (0,1)$, the map $\Psi\colon \mathcal{A} \to \mathbb{R}$ is defined by $\Psi(\mu) := \mathbb{E}_\mu[X]$, and the set of admissible priors π on $\mathcal{A}$ is the collection

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\mu\sim\pi}\left[\mathbb{E}_\mu[X]\right] = q\}\,.$$

for some fixed $q \in (0,a)$. We will now demonstrate how the result $\mathcal{U}(\Pi) = q/a$ of (4.14) obtained by the primary reduction follows from the nested reduction theorem. To that end, observe that since $\Psi(\mathcal{A}) = [0,1] \subseteq \mathbb{R}$, by restricting to measures $\mathbb{Q} \in \mathcal{M}(\mathbb{R})$ with support $\operatorname{supp}(\mathbb{Q}) \subseteq [0,1]$, Theorem 4.11 implies that

$$\mathcal{U}(\Pi) = \sup_{\mathbb{Q}\in\mathfrak{Q}} \mathbb{E}_{q'\sim\mathbb{Q}}\left[\sup_{\mu\in\mathcal{M}([0,1])\,:\,\mathbb{E}_\mu[X]=q'} \mu[X \geq a]\right], \tag{4.20}$$

where $\mathfrak{Q}$ is the set of probability measures $\mathbb{Q}$ on $\mathbb{R}$ with support contained in $[0,1]$ such that $\mathbb{E}_{\mathbb{Q}}[Q] = q$. Theorem 4.1 of [86] shows that the inner supremum of $\mu[X \geq a]$ can be achieved by assuming that μ is the weighted sum of two Dirac masses, i.e.

$$\sup_{\substack{\mu \in \mathcal{M}([0,1]) \\ \mathbb{E}_\mu[X]=q'}} \mu[X \geq a] = \sup_{\substack{\alpha, x_1, x_2 \in [0,1] \\ \alpha x_1 + (1-\alpha) x_2 = q'}} (\alpha \delta_{x_1} + (1-\alpha)\delta_{x_2})[X \geq a]. \tag{4.21}$$

For $q' > a$, the supremum in the right-hand side of (4.21) is 1, and for $q' \leq a$, the supremum in the right-hand side of (4.21) is achieved by $x_2 = 0$, $x_1 = a$ and $\alpha = q'/a$, and so we conclude that

$$\sup_{\substack{\mu \in \mathcal{M}([0,1]) \\ \mathbb{E}_\mu[X]=q'}} \mu[X \geq a] = \min\{1, \tfrac{q'}{a}\}.$$

Hence, by identifying the measures $\mathbb{Q} \in \mathcal{M}(\mathbb{R})$ with support $\mathrm{supp}(\mathbb{Q}) \subseteq [0,1]$ with $\mathcal{M}([0,1])$ in the obvious way, (4.20) becomes

$$\mathcal{U}(\Pi) = \sup_{\substack{\mathbb{Q} \in \mathcal{M}([0,1]) \\ \mathbb{E}_{\mathbb{Q}}[Q]=q}} \mathbb{E}_{q' \sim \mathbb{Q}} \left[\min\{1, \tfrac{q'}{a}\} \right]. \tag{4.22}$$

Using [86, Thm. 4.1] again, we obtain that the supremum in $\mathbb{Q}$ in the right-hand side of (4.22) is equal to the supremum over $\alpha, q_1, q_2 \in [0,1]$, of

$$\alpha \min\{1, \tfrac{q_1}{a}\} + (1-\alpha) \min\{1, \tfrac{q_2}{a}\} \tag{4.23}$$

subject to the constraint that $\alpha q_1 + (1-\alpha) q_2 = q$. This supremum is achieved by $q_1 = a$, $q_2 = 0$ and $\alpha = \frac{q}{a}$, and so we obtain that $\mathcal{U}(\Pi) = q/a$, in agreement with (4.14).

5 Optimal bounds on the posterior value

What happens to the optimal bounds (4.2) and (4.3) on the prior value $\mathbb{E}_\pi[\Phi]$, investigated in Section 4, after conditioning on the data? Does the interval corresponding to these optimal bounds shrink down to a single point as more and more data comes in? Does this interval shrink as the measurement noise on the data is reduced? What happens to posterior estimates associated with two distinct but close priors, possibly sharing the same marginal distribution on a high dimensional space? These are the questions that will be investigated in this section. Our answers will show that: (1) optimal bounds on posterior estimates *grow* as data comes in; (2) optimal bounds on posterior estimates *grow* as measurement noise is reduced (3) two priors sharing the same high-dimensional marginals can lead to *diametrically opposed* posterior estimates. Although the Bayesian framework is a standard method for constructing a statistical estimator, the surprising answers provided in this section will show that one should be cautious with its direct application to UQ.

As discussed in Section 3.5, let us now consider the case where, for a fixed $(f, \mu) \in \mathcal{A}$, there is some uncertainty regarding the observation process $\mathbb{D}(f, \mu)$. Instead of representing this uncertainty by generalizing from $\mathbb{D}\colon \mathcal{A} \to \mathcal{D}$ being a function to $\mathbb{D}\colon \mathcal{A} \twoheadrightarrow \mathcal{D}$ being a set-valued map, we have, for simplicity, chosen to express this uncertainty by specifying a set $\mathfrak{D}$ of mappings from $\mathcal{A}$ to $\mathcal{D}$ and to express our information regarding $\mathbb{D}$ through the assumption $\mathbb{D} \in \mathfrak{D}$. We can easily generalize the notation of Section 3.3 to this more general situation as follows: for an admissible set $\Pi \subseteq \mathcal{M}(\mathcal{A})$ of priors, and a set $\mathfrak{D}$ of observation maps, let $\Pi \odot \mathfrak{D}$ be the set of probability distributions $\pi \odot \mathbb{D}$ on $\mathcal{A} \times \mathcal{D}$ generated by $\pi \in \Pi$ and $\mathbb{D} \in \mathfrak{D}$:

$$\Pi \odot \mathfrak{D} := \{\pi \odot \mathbb{D} \mid \pi \in \Pi,\, \mathbb{D} \in \mathfrak{D}\}. \tag{5.1}$$

As shown in Section 3, directly conditioning measures $\pi \odot \mathbb{D}$ with respect to the random variable D representing the observed sample data would require manipulating regular conditional probabilities on $\mathcal{A} \times \mathcal{D}$.

Furthermore, in Bayesian statistics a prior π may represent a "subjective belief" about reality and, in such situations, the data may be sampled from $\pi^\dagger \cdot \mathbb{D}^\dagger$ which may be distinct from $\pi \cdot \mathbb{D}$. In Bayesian statistics $\pi^\dagger$ is called the "true" (or sometimes "objective") prior and π a "subjective" prior (see [18] and references therein). Although it is known that the subjective prior π might be distinct from the true prior $\pi^\dagger$, one may still try to evaluate the conditional expectation of the quantity of interest Φ using π as the distribution on $\mathcal{A}$. We will show here that although the observation of the sample data d does not uniquely determine the true prior $\pi^\dagger$ and the true data map $\mathbb{D}^\dagger$, it does determine a random subset of $\mathcal{M}(\mathcal{A}) \times \mathfrak{D}$ (i.e. a random subset of priors and data maps) denoted $\mathcal{R}(d)$ such that, with $\pi^\dagger \cdot \mathbb{D}^\dagger$ probability one, $(\pi^\dagger, \mathbb{D}^\dagger) \in \mathcal{R}(d)$. This observation is based on the following fundamental lemma:

Lemma 5.1. *For a strongly Lindelöf space $\mathcal{Y}$ and a Borel measure ν on $\mathcal{B}(\mathcal{Y})$, define*

$$E := \left\{ y \in \mathcal{Y} \,\middle|\, \begin{array}{c} \text{there is an open neighborhood } \mathcal{O}_y \\ \text{of } y \text{ such that } \nu(\mathcal{O}_y) = 0 \end{array} \right\}.$$

Then $\nu(E) = 0$

Remark 5.2. Recall that a Lindelöf space is a topological space such that any open cover has a countable subcover and a strongly Lindelöf space is such that any open subset is Lindelöf. Since $\mathcal{D}$ is assumed to be Suslin from Section 3.3, and Suslin implies strongly Lindelöf, Lemma 5.1 shows that any open neighborhood B_d of any observed value $d \in \mathcal{D}$ has nonzero measure with probability 1.

Remark 5.3. Any separable Hilbert space, in particular the Euclidian space $\mathbb{R}^k$, is strongly Lindelöf. In this situation, Lemma 5.1 implies that if for any observation y generated by a law $\nu \in \mathcal{M}(\mathcal{Y})$ we place an open ball $B(y, r(y))$ of non-zero radius $r(y) > 0$ about y, then with ν-probability 1 we have $\nu\big(B(y, r(y))\big) > 0$. That is,

$$\nu\big(\{y \in \mathcal{Y} \mid \nu(B(y, r(y))) > 0\}\big) = 1.$$

Now suppose the data d are generated according to a probability measure $\pi^\dagger \cdot \mathbb{D}^\dagger$ (where $\pi^\dagger$ is the "true" prior and $\mathbb{D}^\dagger$ the "true" data map). We conclude from Lemma 5.1 that when we observe a sample d, if we assume that $(\pi^\dagger, \mathbb{D}^\dagger) \in \mathcal{R}(d)$ where

$$\mathcal{R}(d) := \left\{(\pi, \mathbb{D}) \in \mathcal{M}(\mathcal{A}) \times \mathfrak{D} \,\middle|\, \pi \cdot \mathbb{D}[B] > 0 \text{ for all } B \text{ open containing } d\right\},$$

then we will be correct in this assumption with $\pi^\dagger \cdot \mathbb{D}^\dagger$-probability 1. Therefore, when the data d are generated and we observe that $d \in B_d$ where B_d is an open subset containing the data d (to keep our notation simple, we will, later on, drop d in the notation B_d), then we restrict our attention to priors $\pi \in \Pi$ and data maps $\mathbb{D} \in \mathfrak{D}$ such that $\pi \cdot \mathbb{D}[B_d] > 0$. That is to say, we restrict our attention to the intersection of $\Pi \odot \mathfrak{D}$ with the set of measures $\pi \odot \mathbb{D}$ such that $(\pi, \mathbb{D}) \in \mathcal{M}(\mathcal{A}) \times \mathfrak{D}$ and $\pi \cdot \mathbb{D}[B_d] > 0$. We write $\Pi \odot_{B_d} \mathfrak{D}$ for this intersection, i.e.

$$\Pi \odot_{B_d} \mathfrak{D} := \{\pi \odot \mathbb{D} \,|\, (\pi, \mathbb{D}) \in \Pi \times \mathfrak{D} \text{ and } \pi \cdot \mathbb{D}[B_d] > 0\}.$$

If $\Pi \odot_{B_d} \mathfrak{D}$ is void, then we assert that "$\pi^\dagger \odot \mathbb{D}^\dagger$ is not contained in $\Pi \odot \mathfrak{D}$" and we know that this assertion is true with $\pi^\dagger \cdot \mathbb{D}^\dagger$-probability 1 on the realization of the data d. Conversely, if $\pi^\dagger \odot \mathbb{D}^\dagger$ is contained in $\Pi \odot \mathfrak{D}$, then $\Pi \odot_{B_d} \mathfrak{D}$ must, with $\pi^\dagger \cdot \mathbb{D}^\dagger$-probability 1 on the realization of the data d, still contain $\pi^\dagger \odot \mathbb{D}^\dagger$ (in particular it must be non-empty).

Happily, this approach also facilitates the efficient computation of the conditional expectations because now they have a simple representation. Indeed, consider the conditional expectation of an object of interest Φ given a prior π and data map $\mathbb{D}$, conditioned on a subset $B \in \mathcal{B}(\mathcal{D})$ such that $\pi \cdot \mathbb{D}[B] > 0$. It follows from (3.4) and (3.5) that the conditional expectation of Φ given B is

$$\mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] := \frac{\mathbb{E}_{(f,\mu,d) \sim \pi \odot \mathbb{D}}\big[\Phi(f,\mu) \mathbb{1}_B(d)\big]}{\pi \cdot \mathbb{D}[B]},$$

which, using (3.4) and (3.5), leads to

$$\mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] = \frac{\mathbb{E}_{(f,\mu) \sim \pi}\big[\Phi(f,\mu) \mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu) \sim \pi}\big[\mathbb{D}(f,\mu)[B]\big]}. \tag{5.2}$$

Moreover, recall that this conditional expectation is the best mean squared approximation of Φ under the measure $\pi \odot \mathbb{D}$, given the information that $D \in B$, i.e.

$$\mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] = \arg\min_{m \in \mathbb{R}} \mathbb{E}_{\pi \odot \mathbb{D}}\Big[\big(\Phi - m\big)^2 \Big| B\Big]. \tag{5.3}$$

Consequently, for any open subset $B \subseteq \mathcal{D}$, we define

$$\Pi \odot_B \mathfrak{D} := \{\pi \odot \mathbb{D} \in \Pi \odot \mathfrak{D} \,|\, (\pi \cdot \mathbb{D})[B] > 0\}. \tag{5.4}$$

where, by (3.5),

$$\pi \cdot \mathbb{D}[B] := \mathbb{E}_{(f,\mu) \sim \pi}\big[\mathbb{D}(f,\mu)[B]\big]. \tag{5.5}$$

Then, since $(\pi \cdot \mathbb{D})[B] > 0$, the formula (5.2) for conditional expectation implies that

$$\mathcal{U}(\Pi \odot_B \mathfrak{D}) := \sup_{\pi \odot \mathbb{D} \in \Pi \odot_B \mathfrak{D}} \mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] \tag{5.6}$$

$$\mathcal{L}(\Pi \odot_B \mathfrak{D}) := \inf_{\pi \odot \mathbb{D} \in \Pi \odot_B \mathfrak{D}} \mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] \tag{5.7}$$

where

$$\mathbb{E}_{\pi \odot \mathbb{D}}\big[\Phi \big| B\big] = \frac{\mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]}. \tag{5.8}$$

Finally, if B is an open neighborhood containing the sample data d, then it follows that $\mathcal{U}(\Pi \odot_{B_d} \mathfrak{D})$ and $\mathcal{L}(\Pi \odot_{B_d} \mathfrak{D})$ are optimal upper and lower bounds on the posterior values $\mathbb{E}_{\pi \odot \mathbb{D}}[\Phi|B]$, given the observation $D \in B$, over all $\pi \in \Pi$ and $\mathbb{D} \in \mathfrak{D}$ such that $\pi \cdot \mathbb{D}[B] > 0$.

Example 5.4. When Φ is the indicator function of the set $\{(f,\mu) \mid \mu[f \geq a] > \epsilon\}$ (i.e. the set of unsafe (f,μ)), $\mathcal{U}(\Pi \odot_B \mathfrak{D})$ and $\mathcal{L}(\Pi \odot_B \mathfrak{D})$ are optimal upper and lower bounds on the "posterior probability" that the system is unsafe given the observation $D \in B$ (and the sets Π and $\mathfrak{D}$ of priors and observation maps respectively).

5.1 General information barriers on posterior values

Now let $B \subseteq \mathcal{D}$ be open and let

$$\mathcal{A}_{\Pi \odot_B \mathfrak{D}} := \left\{ (f,\mu) \in \mathcal{A} \,\middle|\, \delta_{(f,\mu)} \in \Pi \text{ and } \sup_{\mathbb{D} \in \mathfrak{D}} \mathbb{D}(f,\mu)[B] > 0 \right\}, \tag{5.9}$$

$$\mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) := \sup_{(f,\mu) \in \mathcal{A}_{\Pi \odot_B \mathfrak{D}}} \Phi(f,\mu)\,,$$

and use $\mathcal{L}$ for the corresponding infimum. The following theorem is a straightforward consequence of (5.2):

Theorem 5.5. *It holds true that*

$$\mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) \leq \mathcal{U}(\Pi \odot_B \mathfrak{D}) \leq \mathcal{U}(\mathcal{A}),$$

and

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi \odot_B \mathfrak{D}) \leq \mathcal{L}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}).$$

Moreover, if $\mathcal{A}_{\Pi \odot_B \mathfrak{D}}$ is non empty, then

$$\mathcal{L}(\mathcal{A}) \leq \mathcal{L}(\Pi \odot_B \mathfrak{D}) \leq \mathcal{L}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) \leq \mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) \leq \mathcal{U}(\Pi \odot_B \mathfrak{D}) \leq \mathcal{U}(\mathcal{A}).$$

Remark 5.6. The dependence of $\mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}})$ and $\mathcal{L}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}})$ on the sample data is very weak. In particular, if $\mathfrak{D} = \{\mathbb{D}\}$ and $\mathbb{D}$ corresponds to observing i.i.d. realizations of $(X + \xi, f^\dagger(X) + \xi')$ where ξ and ξ' are centered Gaussian random variables of arbitrarily small (non zero) variance, then it can be shown that $\mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) = \mathcal{U}(\mathcal{A}_\Pi)$ and $\mathcal{L}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) =$

$\mathcal{L}(\mathcal{A}_\Pi)$. In that situation, if $\mathcal{L}(\mathcal{A}_\Pi) < \mathcal{U}(\mathcal{A}_\Pi)$, then $\mathcal{U}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}}) - \mathcal{L}(\mathcal{A}_{\Pi \odot_B \mathfrak{D}})$ remains bounded away from 0 by a strictly positive constant that is independent of $\mathfrak{D}$ and B, which, in particular, implies that the range of achievable posterior values cannot shrink towards $\Phi(f^\dagger, \mu^\dagger)$ regardless of the number of observed i.i.d. samples. The presence of such information barriers suggest that the consistency of Bayesian estimators cannot be established independently (uniformly) in the choice of priors (this point will also be substantiated by Theorem 5.12).

5.2 Primary reduction for posterior values

As in Section 4.2.1, when priors are specified through finite-dimensional inequalities, it is possible to provide a reduction of the computation of $\mathcal{U}(\Pi \odot_B \mathfrak{D})$ on the primary space. To that end, let $\mathcal{M}_+(\mathcal{A})$ denote the set of postive bounded measures on $\mathcal{A}$ and let us extend the "expectation notation" to mean integration with respect to a positive measure in the natural way: for a measurable function ψ and a $\pi_+ \in \mathcal{M}_+(\mathcal{A})$ define

$$\mathbb{E}_{\pi_+}[\psi] := \int_{\mathcal{A}} \psi \, \mathrm{d}\pi_+$$

if the integral exists.

Let $\psi_0, \ldots, \psi_n$ be real-valued measurable functions on $\mathcal{A}$ and define

$$\Pi_+ := \left\{ \pi_+ \in \mathcal{M}_+(\mathcal{A}) \,\middle|\, \mathbb{E}_{\pi_+}[\psi_0] = 1, \text{ and } \mathbb{E}_{\pi_+}[\psi_i] = 0 \text{ for } i = 1, \ldots, n \right\},$$

where implicit in the definition is that all $n+1$ integrals exist, and let

$$\Pi_{+,n} := \Pi_+ \cap \Delta(n)$$

be the set of those measures in Π_+ that are non-negative sums of $n+1$ Dirac masses. The following theorem is a generalization of [86, Thm. 4.1] to positive measures (see also [121, Thm. 3.2] from which the proof of [86, Thm. 4.1] was derived).

Theorem 5.7. *If $\mathcal{A}$ is a Suslin space, then*

$$\sup_{\pi_+ \in \Pi_+} \mathbb{E}_{\pi_+}[\Phi] = \sup_{\pi_+ \in \Pi_{+,n+1}} \mathbb{E}_{\pi_+}[\Phi]. \tag{5.10}$$

Furthermore, if ψ_0 is non-negative on $\mathcal{A}$ and there exists a measurable function φ such that $\Phi = \psi_0 \varphi$, then

$$\sup_{\pi_+ \in \Pi_+} \mathbb{E}_{\pi_+}[\Phi] = \sup_{\pi_+ \in \Pi_{+,n}} \mathbb{E}_{\pi_+}[\Phi]. \tag{5.11}$$

Theorem 5.7 can be used to produce a primary reduction of $\mathcal{U}\big(\Pi \odot_B \mathcal{D}\big)$ when Π is defined by a finite number of equalities. To state the theorem, recall that, for arbitrary Π, $\mathfrak{D}$ and B, the definition

$$\Pi \odot_B \mathfrak{D} := \{ \pi \odot \mathbb{D} \in \Pi \odot \mathfrak{D} \,|\, \pi \cdot \mathbb{D}[B] > 0 \}$$

of (5.4), where by (5.5)

$$\pi \cdot \mathbb{D}[B] := \mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big];$$

recall also the notation of (5.6)

$$\mathcal{U}(\Pi \odot_B \mathfrak{D}) := \sup_{\pi\odot\mathbb{D}\in\Pi\odot_B\mathfrak{D}} \mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big];$$

and recall the result (5.2) that, for any $\pi \odot \mathbb{D} \in \Pi \odot_B \mathfrak{D}$,

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \frac{\mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]}.$$

The proof of the following theorem is obtained by first proving the theorem for equality constraints $Z = \{q\}$, by observing that $\mathcal{U}\big(\Pi(q)\odot_B\mathbb{D}\big)$ is a fractional optimization problem in π and utilizing the fact that such problems are equivalent to linear problems [31], and then applying Theorem 5.7. To extend the result to the subset $Z \subseteq \mathbb{R}^n$, one uses a layercake approach as in the proof of Theorem 4.6. As in Section 4, the following primary reduction theorem, Theorem 5.8, will be formulated in canonical form and the nested reduction theorem, Theorem 5.10, will be in the general form.

Theorem 5.8. *Let $\mathcal{A}$ be Suslin and let $\Psi\colon \mathcal{A} \to \mathbb{R}^n$ be measurable. For $Z \subseteq \mathbb{R}^n$, let $\Pi(Z) := \{\pi \in \mathcal{M}(\mathcal{A}) \mid \mathbb{E}_\pi[\Psi] \in Z\}$. Then $\mathcal{U}\big(\Pi(Z) \odot_B \mathcal{D}\big)$ is equal to the supremum over $\mathbb{D} \in \mathcal{D}$, $\alpha_i \geq 0$, $q \in Z$ and $(f_i, \mu_i) \in \mathcal{A}$ of*

$$\sum_{i=0}^{n} \alpha_i \Phi(f_i,\mu_i)\mathbb{D}(f_i,\mu_i)[B]$$

subject to the constraints

$$\sum_{i=0}^{n} \alpha_i\big(\Psi(f_i,\mu_i) - q\big) = 0$$

and

$$\sum_{i=0}^{n} \alpha_i \mathbb{D}(f_i,\mu_i)[B] = 1. \tag{5.12}$$

Example 5.9. Consider again Example 4.7 with the admissible set $\mathcal{A} := \mathcal{M}([0,1])$, the quantity of interest $\Phi(\mu) := \mu[X \geq a]$, the map $\Psi(\mu) := \mathbb{E}_\mu[X]$ and the set of admissible priors

$$\Pi := \big\{\pi \in \mathcal{M}(\mathcal{A}) \,\big|\, \mathbb{E}_{\mu\sim\pi}\big[\mathbb{E}_\mu[X]\big] = q\big\}.$$

for some $q \in (0,a)$. We saw in Example 4.7 that $\mathcal{U}(\Pi) = \frac{q}{a}$. Now suppose that we observe the random variable $D := (X_1,\ldots,X_n)$ corresponding to n i.i.d. samples of $\mu^\dagger \in \mathcal{A}$. More precisely, we observe $D \in B$ where $B = B_1 \times \cdots B_n$ and B_i is the ball in $(0,1)$ of center x_i and radius ρ, $x_i \in (0,1)$ and $0 < \rho \ll 1/n$. Let $\mathbb{D}^n$ denote the data

map corresponding to taking n i.i.d. samples, that is, $\mathbb{D}^n(\mu) := \mu \otimes \cdots \otimes \mu$, and observe that $\mathbb{D}^n(\mu)[B] = \prod_{i=1}^n \mu[B_i]$.

Theorem 5.8 implies that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n)$ is equal to the supremum over $\alpha_1, \alpha_2 \geq 0$, $\mu_1, \mu_2 \in \mathcal{A}$ of

$$\alpha_1 \mu_1[X \geq a]\mathbb{D}^n(\mu_1)[B] + \alpha_2 \mu_2[X \geq a]\mathbb{D}^n(\mu_2)[B]$$

subject to the constraints

$$\alpha_1(\mathbb{E}_{\mu_1}[X] - q) + \alpha_2(\mathbb{E}_{\mu_2}[X] - q) = 0,$$

$$\alpha_1 \mathbb{D}^n(\mu_1)[B] + \alpha_2 \mathbb{D}^n(\mu_2)[B] = 1,$$

with $\mathbb{D}^n(\mu)[B] = \prod_{i=1}^n \mu(B_i)$. Introducing slack variables $\beta_{1,i} := \mu_1[B_i]$ and $\beta_{2,i} := \mu_2[B_i]$ as n linear constraints on μ_1 and n linear constraints on μ_2 we obtain (from [86, Thm. 4.1]) that the supremum can be achieved by assuming that each μ_i is the weighted sum of at most $n+2$ Dirac masses. Assuming that the B_i are non intersecting balls of radius $\rho \ll 1/n$ centered on $x_1, \ldots, x_n$, n of these Dirac masses will have to be put at $x_1, \ldots, x_n$; for optimality, the two others will have to be put at 0 and a (with weights p_1 and p_2). Introducing $\gamma_1 = \alpha_1 \mathbb{D}^n(\mu_1)[B]$ and $\gamma_2 = \alpha_2 \mathbb{D}^n(\mu_2)[B]$, it follows that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n)$ is equal (as $\rho \downarrow 0$) to the supremum over $\gamma_1, \gamma_2 \geq 0$, $p_1, p_2 \in [0,1]$ of

$$\gamma_1 p_1 + \gamma_2 p_2$$

subject to the constraints

$$\gamma_1 + \gamma_2 = 1,$$

and

$$\gamma_1 \frac{(ap_1 + \sum_{i=1}^n x_i \beta_{1,i}) - q}{\prod_{i=1}^n \beta_{1,i}} + \gamma_2 \frac{(ap_2 + \sum_{i=1}^n x_i \beta_{2,i}) - q}{\prod_{i=1}^n \beta_{2,i}} = 0.$$

By considering $0 < \beta_{i,j} \ll 1$ it is easy to obtain that $\mathcal{U}(\Pi \odot_B \mathbb{D}^n) = 1$.

5.3 Nested reduction for posterior values

Here, as in Section 4.2, we show how the optimization problems (5.6) and (5.7) can be reduced to nested OUQ optimization problems (i.e. nested problems analogous to (3.1) and (3.2)) when the collection Π of admissible priors is defined by how they push forward by a measurable mapping $\Psi \colon \mathcal{A} \to \mathcal{Q}$. That is, we specify a feature space $\mathcal{Q}$, a measurable map $\Psi \colon \mathcal{A} \to \mathcal{Q}$, a subset $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ and define the admissible set of priors by

$$\Pi := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \mid \Psi\pi \in \mathfrak{Q}\}.$$

As before, we focus on reducing the upper bound

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q} \odot_B \mathfrak{D}\big) := \sup_{\pi \odot \mathbb{D} \in \Psi^{-1}\mathfrak{Q} \odot_B \mathfrak{D}} \mathbb{E}_{\pi \odot \mathbb{D}}[\Phi|B]. \tag{5.13}$$

Theorem 5.10. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ be measurable. Moreover, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ for all $\mathbb{Q} \in \mathfrak{Q}$. Then, for each $\mathbb{Q} \in \mathfrak{Q}$, $\Psi^{-1}\mathbb{Q}$ is non-empty. Moreover, the upper bound $\mathcal{U}\big(\Psi^{-1}\mathfrak{Q} \odot_B \mathfrak{D}\big)$, defined in (5.13), satisfies*

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q} \odot_B \mathfrak{D}\big) = \sup\left\{ \lambda \in \mathbb{R} \,\middle|\, \sup_{\substack{\mathbb{Q}\in\mathfrak{Q}\\ \mathbb{D}\in\mathfrak{D}}} \mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q)} \big(\Phi(f,\mu) - \lambda\big)\mathbb{D}(f,\mu)[B] \right] > 0 \right\}, \tag{5.14}$$

where the expectations on the right-hand side are defined as in (4.17). Finally, the expectation operator on the right-hand side is measure affine in $\mathbb{Q}$, as defined in (4.3).

Remark 5.11. Note that Theorem 5.10 is more general than Theorem 5.8 because its application does not require the assumption that $\Psi^{-1}\mathfrak{Q}$ is defined via generalized moments constraints.

The following theorem is our Main Brittleness Theorem. It shows not only that the right-hand side of the assertion (5.14) of Theorem 5.10 depends on the sample data in a very weak way, but also that under very mild assumptions the observation of this sample data leads to an increase (rather than a decrease) of the least upper bound on the quantity of interest:

Theorem 5.12. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ be measurable. Moreover, let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ for all $\mathbb{Q} \in \mathfrak{Q}$. Suppose that, for all $\delta > 0$, there exists some $\mathbb{Q} \in \mathfrak{Q}$, $\mathbb{D} \in \mathfrak{D}$ such that*

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\inf_{(f,\mu)\in\Psi^{-1}(q)} \mathbb{D}(f,\mu)[B] \right] = 0 \tag{5.15}$$

and

$$\mathbb{P}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q),\, \mathbb{D}(f,\mu)[B]>0} \Phi(f,\mu) > \sup_{(f,\mu)\in\mathcal{A}} \Phi(f,\mu) - \delta \right] > 0. \tag{5.16}$$

Then

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big) = \mathcal{U}(\mathcal{A}). \tag{5.17}$$

Remark 5.13. Note that the convention that $\sup_{x\in A} \varphi(x) = -\infty$ if A is empty implies that, if the assumption (5.16) is satisfied, then there is a measure $\mathbb{Q} \in \mathfrak{Q}$ such that the set of q such that $\mathbb{D}(f,\mu)[B] > 0$ for some $(f,\mu) \in \Psi^{-1}(q)$ has strictly positive $\mathbb{Q}$-measure.

Remark 5.14. Theorem 5.12 states that if there exists $\mathbb{Q} \in \mathfrak{Q}$ putting some mass on a neighborhood of the values q of Ψ where $\sup_{(f,\mu)\in\Psi^{-1}(q)} \Phi(f,\mu)$ achieves its supremum, then

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big) = \mathcal{U}(\mathcal{A}).$$

On the other hand, Theorem 4.2 asserts that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \leq \mathcal{U}(\mathcal{A}) \tag{5.18}$$

so we conclude that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \leq \mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big), \tag{5.19}$$

That is, *observing the sample data does not improve the optimal bound!* Moreover, when the inequality (5.18) is strict, if we define

$$\delta := \mathcal{U}(\mathcal{A}) - \mathcal{U}(\Psi^{-1}\mathfrak{Q}) > 0$$

then it follows that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) + \delta \leq \mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big), \tag{5.20}$$

from which we conclude that when the inequality (5.18) is strict, *observing the sample data makes the optimal bound worse!* In other words, after the observation of the sample data (which may be limited to a single realization of $(X, f^\dagger(X))$ under the measure $\mu^\dagger$, or an arbitrary large number of independent samples of $(X_i, f^\dagger(X_i))$ the optimal upper bound on the quantity of interest

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) = \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\big]$$

increases to

$$\mathcal{U}(\mathcal{A}) = \sup_{(f,\mu)\in\mathcal{A}} \Phi(f,\mu).$$

Example 5.15. Consider $\mathcal{A} := \mathcal{M}([0,1])$, $\Phi(\mu) = \mathbb{E}_\mu[X]$, $\mathbb{D}^n(\mu) := \mu \otimes \cdots \otimes \mu$. In this example are interested in estimating the mean of X under some unknown measure $\mu^\dagger \in \mathcal{A}$ and we observe $d = (d_1, \ldots, d_n)$, n i.i.d. samples from X; note that n can be very large. The sample data contain information on $\mu^\dagger$ through the fact that their distribution is $\mathbb{D}^n(\mu^\dagger) = \mu^\dagger \otimes \cdots \otimes \mu^\dagger$ (i.e. although the distribution of the sample data is unknown, its dependency structure, as a functional of $\mu^\dagger$, is known).

Let k be a (possibly large) number. Define Π to be the set of priors π under which the distribution of $(\mathbb{E}_\mu[X], \ldots, \mathbb{E}_\mu[X^k])$ is $\mathbb{Q}$, where $\mathbb{Q}$ is a distribution on $\mathbb{R}^k$ such that $\mathbb{E}_\mu[X]$ (its first marginal) is uniformly distributed on $[0,1]$ and such that the (conditional) distribution of $\mathbb{E}_\mu[X^2]$ conditioned on $\mathbb{E}_\mu[X] = q_1$ is the uniform distribution on the interval

$$\left[\inf_{\mu\in\mathcal{A},\, \mathbb{E}[X]=q_1} \mathbb{E}_\mu[X^2], \sup_{\mu\in\mathcal{A},\, \mathbb{E}[X]=q_1} \mathbb{E}_\mu[X^2]\right]$$

and such that the conditional distributions of the other marginals $\mathbb{E}_\mu[X^k]$ are defined iteratively in the same manner. For this example, note that $\Psi(\mu) = (\mathbb{E}_\mu[X], \ldots, \mathbb{E}_\mu[X^k])$. Note that, for $q := (q_1, \ldots, q^k)$ in the range of Ψ (i.e. $\Psi(\mathcal{A})$), $\Psi^{-1}(q)$ is the subset of measures $\mu \in \mathcal{M}([0,1])$ such that $\mathbb{E}_\mu[X^i] = q_i$ for $1 \leq i \leq k$. Let B be defined as $B_1 \times \cdots B_n$ where each B_i is a ball of radius ρ containing d_i.

We will now use Theorem 5.12 to compute optimal bounds on the posterior values of $\Phi(\mu) = \mathbb{E}_\mu[X]$. We will focus our attention on the upper bound. First observe that in this example $\mathfrak{Q}$ is reduced to the single measure $\mathbb{Q}$ constructed above and $\mathfrak{D}$ is reduced to the single data map $\mathbb{D}^n$.

Let us first check that condition (5.16) is always satisfied (irrespective of the value of the data d). Note that condition (5.16) is satisfied if for all $\delta > 0$ there exists a subset of values of q of strictly positive $\mathbb{Q}$-measure such that $\{\mu \in \Psi^{-1}(q) \mid \mathbb{D}^n(\mu)[B] > 0 \text{ and } \mathbb{E}_\mu[X] \geq 1 - \delta\}$ is non empty. So, let $\delta > 0$ be arbitrary and define μ_d to be the empirical distribution of d, i.e.

$$\mu_d := \frac{\sum_{i=1}^n \delta_{d_i}}{n}$$

Define

$$\mathcal{A}_\delta := \{\mu \in \mathcal{A} \mid \mathbb{E}_\mu[X] \geq 1 - \delta/2\}.$$

One can show by induction that $\Psi(\mathcal{A}_\delta)$ has a non-empty interior and that any open subset of $\Psi(\mathcal{A})$ has strictly positive $\mathbb{Q}$-measure. Let q^* be a point in the interior of $\Psi(\mathcal{A}_\delta)$, and let $B(q^*, \tau)$ be a ball of center q^* and radius τ such that $B(q^*, 2\tau)$ is contained in the interior of $\Psi(\mathcal{A}_\delta)$. Note that $B(q^*, \tau)$ has strictly positive $\mathbb{Q}$-measure. Furthermore, for ϵ sufficiently small, for each $q \in B(q^*, \tau)$ there exists $q' \in B(q^*, 2\tau)$ and $\mu \in \Psi^{-1}(q')$ such that $\mu_\epsilon := (1-\epsilon)\mu + \epsilon\mu_d \in \Psi^{-1}(q)$. Since $\mathbb{D}^n(\mu_\epsilon)[B] > 0$ and $\mathbb{E}_\mu[X] \geq 1 - \delta/2$, it follows that (5.16) is satisfied (irrespective of the value of the data d).

Let us now consider condition (5.15). Observe that condition (5.15) is satisfied if for $\mathbb{Q}$-almost all $q \in \Psi(\mathcal{A})$ and all $\epsilon > 0$, there exists $\mu \in \Psi^{-1}(q)$ such that $\mathbb{D}^n(\mu)[B] < \epsilon$. Assume that d contains at least $k+2$ distinct points and that ρ is strictly smaller than half of the minimal distance between two of such points, so that the associated B_i do not overlap; note that this assumption is satisfied with probability converging to one (as $n \to \infty$) if the data are sampled from a measure $\mu^\dagger$ that is absolutely continuous with respect to the Lebesgue measure on $[0,1]$. Let $q \in \Psi(\mathcal{A})$; by the reduction theorems of [86] there exists $\mu_q \in \Psi^{-1}(q)$ such that μ_q is the weighted sum of at most $k+1$ masses of Diracs on $[0,1]$. Since there exist at least $k+2$ non-overlapping B_i we have $\mathbb{D}^n(\mu_q)[B] = 0$ which implies condition (5.15). Hence, Theorem 5.12 implies that, for this (possibly) highly constrained problem characterized by a (possibly) large number of sampled data points, the optimal bounds on the posterior values of $\mathbb{E}_\mu[X]$ are zero and one whereas the set of prior values of $\mathbb{E}_\mu[X]$ is the single point $\{\frac{1}{2}\}$.

Remark 5.16. For a thorough analysis of Example 5.15 we refer to [85] where, in particular, a *quantitative* version of Theorem 5.12 is developed and then applied to Example 5.15. This application also leads to the discovery of a new family of Selberg integral formulas through a refined analysis of the integral geometry of the Hausdorff moment space through the revelation that the free parameter associated with Markov and Kreĭn's canonical representations of truncated Hausdorff moments generates reproducing kernel identities corresponding to reproducing kernel Hilbert spaces of polynomials.

Remark 5.17. Note that the assumptions of Theorem 5.12 are extremely weak. In plain words, Theorem 5.12 implies that if the probability of observing the data can be

arbitrary small under priors contained in $\mathcal{A}$ that are putting mass near the extreme values of Φ, then the optimal bounds on posterior values are the extreme values of Φ in $\mathcal{A}$ (even if the data comes in the form of a large number of samples and the set of priors is highly constrained). Example 5.15 illustrates that one consequence of Theorem 5.12 is that Bayesian posteriors are not robust, and in fact are fragile with respect to the choices of priors constrained by marginals, even with a highly constrained subset of priors of $\mathcal{M}(\mathcal{A})$. Moreover, if Π is convex, then by considering priors of the form $\pi_0\lambda + (1-\lambda)\pi_1$ with $\pi_0, \pi_1 \in \Pi$, $\pi_0 \cdot \mathbb{D}[B] > 0$ and $\pi_1 \cdot \mathbb{D}[B] > 0$, it is easy to see that the Bayesian posterior can be "anything you want" in the interval $\big(\mathcal{L}(\mathcal{A}), \mathcal{U}(\mathcal{A})\big)$ (irrespective of the data) in the sense that, for any value I in that interval, there exists a prior $\pi \in \Psi^{-1}(\mathfrak{Q})$ whose posterior value is I. In addition, it is easy to observe that including the quantity of interest Φ in the marginal Ψ does not prevent this fragility. Theorem 5.12 also leads to the following apparent paradoxes when the Bayesian framework is applied to the space $\mathcal{A}$: (1) Posteriors with different priors may diverge as more and more data comes in; (2) When the sample data is observed with some (say Gaussian) measurement noise of variance σ^2, then, writing $\mathbb{D}(\sigma^2)$ for the associated measurement map, the optimal bound $\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathbb{D}(\sigma^2)\big)$ on the quantity of interest Φ converges towards $\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\big)$ as $\sigma^2 \to \infty$. That is, if one interprets optimal bounds on posterior values as uncertainty bounds, then one would reach the paradoxical conclusion that adding measurement uncertainty decreases the uncertainty of the quantity of interest. The idea of the proof of this assertion is based on the following observation:

Let y be the (noisy) measurement whose distribution given the value of the data d is assumed to be independent of (f, μ). Write $p_\sigma(d)[B]$ for the probability that the value of y belongs to a set B and observe that the conditional value of the quantity of interest Φ given the $y \in B$ is equal to

$$\frac{\mathbb{E}_\pi\Big[\Phi(f,\mu)\mathbb{E}_{d\sim\mathbb{D}(f,\mu)}\big[p_\sigma(d)[B]\big]\Big]}{\mathbb{E}_\pi\Big[\mathbb{E}_{d\sim\mathbb{D}(f,\mu)}\big[p_\sigma(d)[B]\big]\Big]}. \tag{5.21}$$

We deduce that if $p_\sigma(d)[B]/p_\sigma(d')[B]$ converges towards one as the level of noise $\sigma \to \infty$ uniformly in $(d, d') \in [0,1]^2$ (which is the case if the data in Example 5.9 is observed with Gaussian noise of increasing variance, see also Example 5.18 below), then (5.21) converges towards the prior value of Φ as $\sigma \to \infty$ uniformly in π.

Example 5.18. Consider again Example 5.9 with the set of admissible priors π on $\mathcal{A}$ defined as the collection

$$\Pi := \big\{\pi \in \mathcal{M}(\mathcal{A}) \,\big|\, \mathbb{E}_{\mu\sim\pi}\big[\mathbb{E}_\mu[X]\big] = q\big\}.$$

and the map $\mathbb{D}^n$ corresponding to the observation of n i.i.d. samples of μ. For $q \in (0, a)$, let $\mathfrak{Q}$ be the set of probability measures $\mathbb{Q}$ on $[0,1]$ such that $\mathbb{E}_{q'\sim\mathbb{Q}}[q'] = q$. Let $\mathbb{Q}$ be the probability measure on $[0,1]$ with probability density function $p(x) = (1-q)/q$ on $[0,q]$ and $p(x) = q/(1-q)$ on $(q,1]$. It is easy to check that $\mathbb{Q} \in \mathfrak{Q}$, that

$$\mathbb{E}_{q'\sim\mathbb{Q}}\left[\inf_{\mu\in\mathcal{A}\,:\,\mathbb{E}_\mu[X]=q'} \prod_{i=1}^n \mu[B_i]\right] = 0, \tag{5.22}$$

and that, for all $\delta > 0$,

$$\mathbb{P}_{q' \sim \mathbb{Q}} \left[\sup_{\mu \in \mathcal{A} : \mathbb{E}_\mu[X] = q', \prod_{i=1}^n \mu[B_i] > 0} \mathbb{E}_\mu[X] > 1 - \delta \right] > 0. \tag{5.23}$$

It follows from Theorem 5.12 that

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big) = 1. \tag{5.24}$$

Remark 5.19. It is known from the Bernstein–von Mises theorem [24, 110] that, in finite-dimensional situations, posterior values converge towards the quantity of interest if the prior distribution has strictly positive mass in every neighbourhood of the truth (see also [76, 84]). It is also known that "even for the simplest infinite-dimensional models, the Bernstein-von Mises theorem does not hold" [37, 53]. This possible lack of convergence, referred to as the consistency problem, has been at the center of a debate between frequentists and Bayesians. We quote Diaconis and Freedman [42] (see also [43])

> "If the underlying mechanism allows an infinite number of possible outcomes (e.g., estimation of an unknown probability on the integers), Bayes estimates can be inconsistent: as more and more data comes in, some Bayesian statisticians will become more and more convinced of the wrong answer."

What is the significance of Theorem 5.12 in that discussion? To answer this question, consider Example 5.9 (and 5.18), in which one is interested in estimating the probability (under the unknown measure $\mu^\dagger$) that X exceeds a after observating n independent samples. We already know from [42, 37] that placing priors on the infinite dimensional space $\mathcal{A} = \mathcal{M}[0,1]$ of probability measures on $[0,1]$ is unlikely to lead to Bayesian posteriors that will converge towards the true value as more and more data comes in. One strategy to circumvent this lack of convergence would be to consider a finite-dimensional subset of $\mathcal{A}$, i.e. a family (μ_λ) of probability measures on $[0,1]$ indexed by a finite-dimensional parameter $\lambda \in \mathbb{R}^k$, put a strictly positive prior p on $\lambda \in \mathbb{R}^k$, and then invoke the Bernstein–von Mises theorem to guarantee the convergence of posterior values.

However, the Bernstein–von Mises theorem requires that the true distribution under which the data is sampled belongs to $\{\mu_\lambda \mid \lambda \in \mathbb{R}^k\}$, the parametrized finite-dimensional subset of $\mathcal{A}$. What happens when this is not the case, i.e. the situation of *misspecification*? Write π_p for the push-forward of the prior p on $\lambda \in \mathbb{R}^k$ to a prior on $\mathcal{A}$ under the map $\lambda \mapsto \mu_\lambda$. Assume that $\mathfrak{D} = \{\mathbb{D}\}$ and that the data have been sampled from $\pi^\dagger \cdot \mathbb{D}$ where $\pi^\dagger$ is the (frequentist) true distribution. Here Theorem 5.12, as illustrated in Example 5.15, can be used to show that the posterior values of the quantity of interest under π_p and $\pi^\dagger$ may lie near the opposite extreme values of Φ in $\mathcal{A}$ even if (1) $\pi^\dagger$ is a Dirac mass on a measure $\mu^\dagger \in \mathcal{A}$; (2) the number of independent samples is large; and (3) k is large and k moments of $\mu^\dagger$ and μ_{λ^*} are equal for some $\lambda^* \in \mathbb{R}^k$.

5.4 Min-Max Bayesian posterior

Equation (5.3) implies that

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}\big)=\sup_{\pi\odot\mathbb{D}\in\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}}\ \arg\min_{m\in\mathbb{R}}\mathbb{E}_{\pi\odot\mathbb{D}}\Big[\big(\Phi-m\big)^2\Big|B\Big] \tag{5.25}$$

and Theorem 5.12 shows that, under very general conditions, $\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}\big)=\mathcal{U}(\mathcal{A})$, which implies the fragility of the Bayesian posterior with respect to the choice of the prior. It is natural to wonder whether this fragility can be remediated by using a min-max version of the conditional expectation defined by switching the positions of the supremum in π with that of the minimum in m in (5.25). More precisely, using the notation of Section 5.3, we define the *min-max conditional expectation* as

$$\arg\min_{m\in\mathbb{R}}\ \sup_{\pi\odot\mathbb{D}\in\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}}\mathbb{E}_{\pi\odot\mathbb{D}}\Big[\big(\Phi-m\big)^2\Big|B\Big] \tag{5.26}$$

Although we have not found previous references to our version of a min-max Bayesian posterior value, our definition is motivated by:

1. The observation that min-max definitions have previously been employed in making decisions or predictions under uncertainty. We refer, for instance to the introduction of the worst-case conditional Value-at-Risk and its application to robust portfolio management [123] and to robust min-max portfolio strategies for rival forecast and risk scenarios [93].

2. The question of whether such definition could resolve the lack of convergence of the posterior value.

The following theorem shows the answer to (2) is *no*, and that, in particular, (5.26) is in general equal to the midpoint of the OUQ interval $\big[\mathcal{L}(\mathcal{A}),\mathcal{U}(\mathcal{A})\big]$, i.e. $\frac{\mathcal{U}(\mathcal{A})+\mathcal{L}(\mathcal{A})}{2}$. Hence, the min-max conditional expectation cannot converge towards $\Phi(f^\dagger,\mu^\dagger)$.

Theorem 5.20. *Let $\mathcal{A}$ be a Suslin space, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon\mathcal{A}\to\mathcal{Q}$ be measurable. Moreover, let $\mathfrak{Q}\subseteq\mathcal{M}(\mathcal{Q})$ be such that* $\operatorname{supp}(\mathbb{Q})\subseteq\Psi(\mathcal{A})$ *for all $\mathbb{Q}\in\mathfrak{Q}$. Suppose that, for all $\delta>0$, there exists $\mathbb{Q}\in\mathfrak{Q}$, $\mathbb{D}\in\mathfrak{D}$, such that*

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\inf_{(f,\mu)\in\Psi^{-1}(q)}\mathbb{D}(f,\mu)[B]\right]=0, \tag{5.27}$$

$$\mathbb{P}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q),\,\mathbb{D}(f,\mu)[B]>0}\Phi(f,\mu)>\sup_{(f,\mu)\in\mathcal{A}}\Phi(f,\mu)-\delta\right]>0, \tag{5.28}$$

and

$$\mathbb{P}_{q\sim\mathbb{Q}}\left[\inf_{(f,\mu)\in\Psi^{-1}(q),\,\mathbb{D}(f,\mu)[B]>0}\Phi(f,\mu)<\inf_{(f,\mu)\in\mathcal{A}}\Phi(f,\mu)+\delta\right]>0. \tag{5.29}$$

Then

$$\arg\min_{m\in\mathbb{R}}\ \sup_{\pi\odot\mathbb{D}\in\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}}\mathbb{E}_{\pi\odot\mathbb{D}}\Big[\big(\Phi-m\big)^2\Big|B\Big]=\frac{\mathcal{U}(\mathcal{A})+\mathcal{L}(\mathcal{A})}{2}.$$

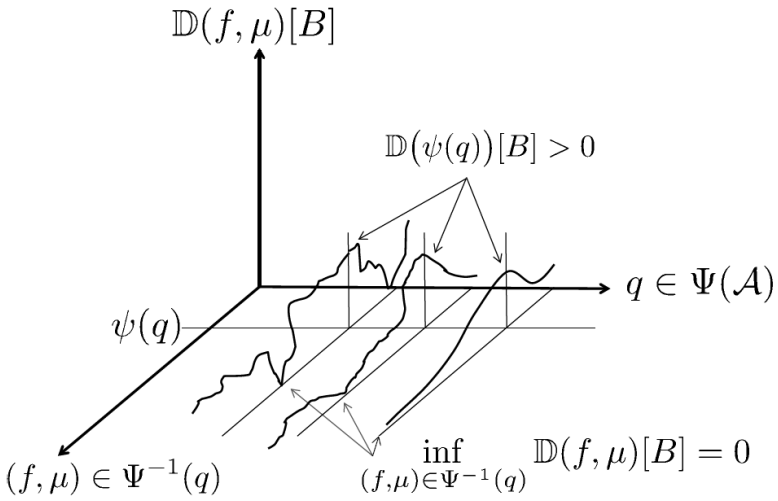

Figure 6.1: Illustration of Conditions (6.1) and (6.2) of Theorem 6.1. If, for some data map $\mathbb{D} \in \mathfrak{D}$, all level sets of Ψ go to zero (i.e. for all $q \in \Psi(\mathcal{A})$, $\inf_{(f,\mu)\in\Psi^{-1}(q)} \mathbb{D}(f,\mu)[B] = 0$), then, for any positive section ψ of Ψ (i.e. $\Psi \circ \psi(q) = q$ and $\mathbb{D}(\psi(q))[B] > 0$ for $q \in \Psi(\mathcal{A})$), the least upper bound on posterior values is bounded from below by the essential supremum of $\Phi \circ \psi$.

6 Brittleness under Local Misspecification

We now establish a corollary to the proof of Theorem 5.12 which we will then use to establish an extreme brittleness theorem for a Bayesian model with local misspecification. Recall that, for a map $\Psi\colon \mathcal{A} \to \mathcal{Q}$, a map $\psi\colon \Psi(\mathcal{A}) \to \mathcal{A}$ is called a *section* of Ψ if $\Psi \circ \psi(q) = q$ for all $q \in \Psi(\mathcal{A})$.

Theorem 6.1. *Let $\mathcal{A}$ be a Suslin space, let $\Phi\colon \mathcal{A} \to \mathbb{R}$ be measurable, let $\mathcal{Q}$ be a separable and metrizable space, and let $\Psi\colon \mathcal{A} \to \mathcal{Q}$ measurable. Let $\mathfrak{Q} \subseteq \mathcal{M}(\mathcal{Q})$ be such that $\operatorname{supp}(\mathbb{Q}) \subseteq \Psi(\mathcal{A})$ for all $\mathbb{Q} \in \mathfrak{Q}$. Let the data space $\mathcal{D}$ be metrizable and consider $B \in \mathcal{B}(\mathcal{D})$. Let $\mathbb{D} \in \mathfrak{D}$ be such that all the level sets of Ψ go to zero, in the sense that*

$$\inf_{(f,\mu)\in\Psi^{-1}(q)} \mathbb{D}(f,\mu)[B] = 0, \quad \text{for all } q \in \Psi(\mathcal{A}). \tag{6.1}$$

Then for any positive measurable section ψ of Ψ, positive in the sense that

$$\mathbb{D}(\psi(q))[B] > 0, \quad \text{for all } q \in \Psi(\mathcal{A}), \tag{6.2}$$

it follows that

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q}) \odot_B \mathfrak{D}\big) \geq \mathfrak{Q}^{\infty}(\Phi \circ \psi). \tag{6.3}$$

where $\mathfrak{Q}^{\infty}(\Phi \circ \psi)$ is the essential supremum

$$\mathfrak{Q}^{\infty}(\Phi \circ \psi) := \sup_{\mathbb{Q}\in\mathfrak{Q}} \inf \big\{ r \in \mathbb{R} : \mathbb{Q}[\Phi \circ \psi > r] = 0 \big\}. \tag{6.4}$$

See Figure 6.1 for an illustration of Theorem 6.1.

We now use Theorem 6.1 to develop a brittleness theorem for a Bayesian model with local misspecification. To that end, let $\mathcal{X}$ be a Polish space so that, by [4, Thm.15.15], $\mathcal{M}(\mathcal{X})$ endowed with the weak-$*$ topology is Polish. Moreover, by [48, Thm. 11.3.3], we know that if we select a complete consistent metric d for $\mathcal{X}$, then the Prokhorov metric $d_{\mathcal{M}}$ defined by

$$d_{\mathcal{M}}(\mu_1, \mu_2) := \inf \{\varepsilon > 0 \,|\, \mu_1(A) \leq \mu_2(A^\varepsilon) + \varepsilon \text{ for all } A \in \mathcal{B}(\mathcal{X})\},$$

where

$$A^\varepsilon := \left\{x \in \mathcal{X} \,\middle|\, d(x, x') < \varepsilon \text{ for some } x' \in A\right\}$$

is the ε neighborhood of A, metrizes the weak-$*$ topology on $\mathcal{M}(\mathcal{X})$. Moreover, Prokhorov's Theorem [48, Cor. 11.5.5] asserts that the Prokhorov metric $d_{\mathcal{M}}$ is a complete metric for the Polish space $\mathcal{M}(\mathcal{X})$. For $\alpha > 0$, $\mu \in \mathcal{M}(\mathcal{X})$, let $B_\alpha(\mu) := \{\mu' \in \mathcal{M}(\mathcal{X}) \mid d_{\mathcal{M}}(\mu, \mu') < \alpha\}$ be the open ball of Prokhorov radius α about μ.

Let Θ be a Polish space and let the Bayesian model define a map

$$\mathcal{P}\colon \Theta \to \mathcal{M}(\mathcal{X}).$$

As in Section 2, the image $\mathcal{P}(\Theta)$ is referred to as the (Bayesian) *model class*.

Remark 6.2. When $\mathcal{P}$ is continuous, it follows from the definition [6, Sec. 3.2] of an analytic set that the the image $\mathcal{P}(\Theta) \subseteq \mathcal{M}(\mathcal{X})$ is analytic, and since the range space $\mathcal{M}(\mathcal{X})$ is Polish it follows that $\mathcal{P}(\Theta)$ is Suslin. Actually, continuity is not required, since [6, Thm. 3.3.4] implies that if $\mathcal{P}$ is measurable, then the image $\mathcal{P}(\Theta)$ is Suslin. If, in addition, $\mathcal{P}$ is injective, then Suslin's Theorem [6, Thm. 3.2.3] implies that $\mathcal{P}(\Theta)$ is Borel.

Assume that $\mathcal{P}$ is measurable and denote its image by $\mathcal{A}_0 := \mathcal{P}(\Theta)$. Since $\mathcal{A}_0 \subseteq \mathcal{M}(\mathcal{X})$ is Suslin, it follows from Lemma 7.2 that the push-forward operator $\mathcal{P}\colon \mathcal{M}(\Theta) \to \mathcal{M}(\mathcal{A}_0)$ is affine continuous. Let $\pi_\Theta \in \mathcal{M}(\Theta)$ be a prior distribution on Θ and let $\pi_0 := \mathcal{P}\pi_\Theta \in \mathcal{M}(\mathcal{A}_0)$ be its pushforward.

Let $\Phi_0\colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ be a measurable quantity of interest. We are interested in estimating Φ_0 using the prior π_0 and our purpose is to show the extreme brittleness of this estimation under arbitrarily small perturbations of the model class $\mathcal{A}_0$ in both the Prokhorov and total variation metrics.

For conditioning on observations, let the data space be $\mathcal{D} := \mathcal{X}^n$, and consider the n-i.i.d. sample data map $\mathbb{D}_0^n\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}^n)$ defined by

$$\mathbb{D}_0^n \mu := \mu^n, \quad \mu \in \mathcal{M}(\mathcal{X}). \tag{6.5}$$

For $x^n = (x_1, \ldots, x_n) \in \mathcal{X}^n$, dropping the notational dependence, denote the rectangle about x^n by

$$B_\delta^n := \prod_{i=1}^n B_\delta(x_i), \tag{6.6}$$

where $B_\delta(x_i)$ is the open ball of radius δ about x_i.

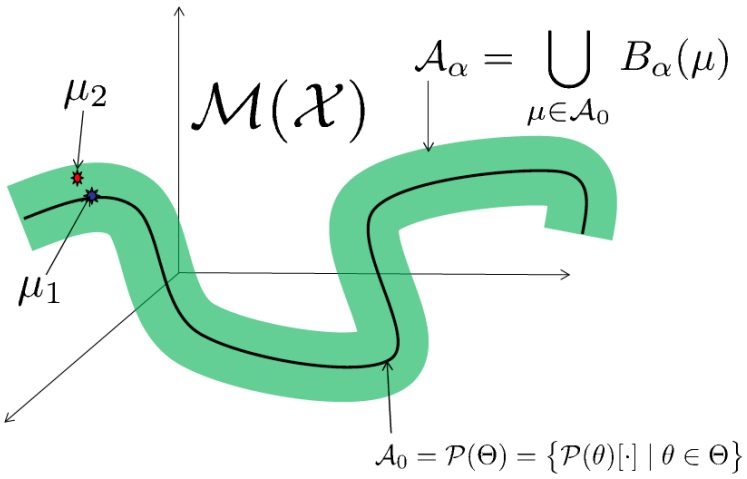

Figure 6.2: Illustration of $\mathcal{A}_0$, $\mathcal{A}_\alpha$ and $(\mu_1, \mu_2) \in \mathcal{A}$.

Observe that the prior value of Φ_0 under π_0 is $\mathbb{E}_{\pi_0}[\Phi_0]$ and its posterior value under the observation $d \in B_\delta^n$ is $\mathbb{E}_{\pi_0 \odot_{B_\delta^n} \mathbb{D}_0^n}[\Phi_0]$.

To define α-perturbations of the model class $\mathcal{A}_0$ in Prokhorov metric, we introduce, for $\alpha > 0$ the α-neighborhood $\mathcal{A}_\alpha \subseteq \mathcal{M}(\mathcal{X})$ of $\mathcal{A}_0$ defined by

$$\mathcal{A}_\alpha := \bigcup_{\mu\in\mathcal{A}_0} B_\alpha(\mu). \tag{6.7}$$

It is easy to see that the ball fibration (see Figure 6.2 and Remark 6.9)

$$\mathcal{A} := \{(\mu_1, \mu_2) \in \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \,|\, \mu_1 \in \mathcal{A}_0, \mu_2 \in B_\alpha(\mu_1)\} \tag{6.8}$$

of the set of balls about points of $\mathcal{A}_0$ projects to

$$P_0\mathcal{A} = \mathcal{A}_0 \tag{6.9}$$
$$P_\alpha\mathcal{A} = \mathcal{A}_\alpha \tag{6.10}$$

where $P_0 \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X})$ is the projection onto the first component and P_α the projection onto the second. The naturally induced set of priors corresponding to $\pi_0 \in \mathcal{M}(\mathcal{A}_0)$ is therefore the set $\Pi_\alpha \subset \mathcal{M}(\mathcal{A}_\alpha)$ defined by

$$\Pi_\alpha := \left\{\pi_\alpha \in \mathcal{M}(\mathcal{A}_\alpha) \middle| \exists \pi \in \mathcal{M}(\mathcal{A}) \text{ with } P_0\pi = \pi_0 \text{ and } P_\alpha\pi = \pi_\alpha\right\}. \tag{6.11}$$

Remark 6.3. Observe that each element $\pi_\alpha \in \Pi_\alpha$ is the distribution of a random measure μ_2 on $\mathcal{A}_\alpha$ such that: (i) there exists a random measure $\mu_1 \in \mathcal{A}_0$ with distribution π_0 (that of the Bayesian model) (ii) (μ_1, μ_2) is jointly-measurable (iii) with probability one the Prokhorov distance from μ_2 to μ_1 is less than α, i.e. $d_{\mathcal{M}}(\mu_1, \mu_2) < \alpha$. Observe in particular that $\pi_0 \in \Pi_\alpha$.

Our main result is provided in Theorem 6.10 but for the sake of clarity we will first give this result in the following (simpler) form.

Theorem 6.4. *Using the notations introduced above, let* Π_α *be defined as in* (6.11)*. If*

$$\lim_{\delta \downarrow 0} \sup_{x \in \mathcal{X}} \sup_{\theta \in \Theta} \mathcal{P}(\theta)[B_\delta(x)] = 0, \tag{6.12}$$

then, for all $\alpha > 0$ *there exists* $\delta_c(\alpha) > 0$ *such that for all* $0 < \delta < \delta_c(\alpha)$ *and all integers* $n \geq 1$*,*

$$\mathcal{U}(\Pi_\alpha \odot_{B_\delta^n} \mathbb{D}_0^n) \geq \operatorname{ess\,sup}_{\pi_0}(\Phi_0)$$

where

$$\operatorname{ess\,sup}_{\pi_0}(\Phi_0) := \inf\{r > 0 \mid \pi_0[\phi_0 > r] = 0\}$$

and with similar expressions for the lower bounds $\mathcal{L}$*.*

Remark 6.5. Theorem 6.4 implies the extreme brittleness of Bayesian inference under local misspecification. Indeed, assume that the model class $\mathcal{A}_0$ is well specified (i.e. it contains the truth $\mu^\dagger$) and that, therefore, the Bayesian estimator described by π_0 is consistent. One may believe that a model $\mathcal{A}_1$ lying in a "small enough" neighborhood of $\mathcal{A}_0$ should have good convergence properties, Theorem 6.4 and Remark 6.3 invalidate this belief. Using the notations of Remark 6.3, observe in particular that an unscrupulous practitioner may design a model corresponding to a random measure μ_2 such that the distance between μ_1 (the well specified model) and μ_2 is a.s. at most α (where α is arbitrarily small) and the posterior value using the random measure μ_2 is as distant as possible from the posterior value using μ_1 irrespective of the sample size n.

Remark 6.6. Observe that the condition (6.12) is extremely weak and satisfied for most Bayesian models. This condition can in fact be made weaker by replacing it with the assumption that for n sufficiently large it holds true that for all θ, $\mathcal{P}(\theta)$ does not contain a mass of Dirac in each ball $B_\delta(x_i)$ (i.e. on the sample data when $\delta \downarrow 0$). We also note that the proof of Theorem 6.4 does not require the samples to be i.i.d., in particular, the same results can be obtained with coupled samples, if, for instance, the data map $\mathbb{D}_0^n$ is replaced by a data map $\mathbb{D}$ such that $C_1^n \prod_{i=1}^n \mu(A_i) \leq \mathbb{D}(\mu)[A_1 \times \cdots \times A_n] \leq C_2^n \prod_{i=1}^n \mu(A_i)$ for strictly positive constants C_1 and C_2.

Remark 6.7. Theorem 6.4 is a corollary of Theorem 6.10 and the proof of Theorem 6.10 shows that, if Θ is compact and $\mathcal{P}$ is continuous and $\Phi(\mu) := \mu(A)$ for some fixed $A \in \mathcal{B}(\mathcal{X})$, then (see Remark 6.13) the result of Theorem 6.4 holds when using the total variation distance d_{TV} instead of the Prokhorov distance, which produces a much smaller neighborhood.

Remark 6.8. Theorems 5.12 and 6.4 are a posteriori brittleness estimates, i.e. posterior to the observation of the data. Note that under the (weak) condition (6.12) the conclusion of Theorem 6.4 holds uniformly irrespective of the size and value of the data. We will show in a sequel work that the brittleness of Bayesian Inference is even stronger

with respect to a priori statistical estimation error estimates (i.e., after averaging with respect to the data generating distribution) because of the possible singularity of that distribution with respect to the model under misspecification.

We will now give a more general version of Theorem 6.4 and elaborate on the objects entering in its formulation.

We start with $\Pi_\Theta \subseteq \mathcal{M}(\Theta)$, a set of admissible priors and let

$$\Pi_0 := \mathcal{P}\Pi_\Theta \subseteq \mathcal{M}(\mathcal{A}_0)$$

denote the push-forward by the model $\mathcal{P}$.

We consider the pull-back $\Phi_\Theta := \Phi_0 \circ \mathcal{P}$, of the measurable quantity of interest $\Phi_0 \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$, to a measurable quantity of interest $\Phi_\Theta \colon \Theta \to \mathbb{R}$. Then the change of variables formula [48, Thm. 4.1.11] implies that, for $\pi_\Theta \in \mathcal{M}(\Theta)$,

$$\mathbb{E}_{\pi_\Theta}[\Phi_\Theta] = \mathbb{E}_{\pi_\Theta}[\Phi_0 \circ \mathcal{P}] = \mathbb{E}_{\mathcal{P}\pi_\Theta}[\Phi_0]$$

whenever either side is well defined. Therefore, taking supremums and infimums, we obtain

$$\begin{aligned}\mathcal{U}(\Pi_\Theta) &= \mathcal{U}(\Pi_0),\\ \mathcal{L}(\Pi_\Theta) &= \mathcal{L}(\Pi_0),\end{aligned}$$

where we note that the quantity of interest implicit in these definitions is determined by the argument. For $\alpha > 0$, define $\mathcal{A}_\alpha$, $\mathcal{A}$, P_0 and P_α as in (6.7), (6.8), (6.9) and (6.10).

Remark 6.9. Using the affine convexity of $\mathcal{M}(\mathcal{X})$, one can show that $\mathcal{A}$ is indeed a Hurewicz fibration, in that it has the homotopy lifting property, see e.g. [99, Pg. 66]. Since $d_{\mathcal{M}} \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ is continuous, it follows that $d_{\mathcal{M}}^{-1}(<\alpha) := \{(\mu_1, \mu_2) \mid d_{\mathcal{M}}(\mu_1, \mu_2) < \alpha\}$ is open and therefore Borel. In addition, since $\mathcal{A}_0 \subseteq \mathcal{M}(\mathcal{X})$ is Suslin it follows that $\mathcal{A}_0 \times \mathcal{M}(\mathcal{X}) \subseteq \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is Suslin. Therefore, since $\mathcal{A} = d_{\mathcal{M}}^{-1}(<\alpha) \cap \big(\mathcal{A}_0 \times \mathcal{M}(\mathcal{X})\big)$ it follows that $\mathcal{A}$ is Suslin.

Observe that the measurable quantity of interest $\Phi_0 \colon \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ acting on the second component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, naturally pulls back to the quantity of interest $\Phi \colon \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathbb{R}$ by $\Phi := \Phi_0 \circ P_\alpha$, and we have $\sup_{\mathcal{A}_\alpha} \Phi_0 = \sup_{\mathcal{A}} \Phi$ and $\inf_{\mathcal{A}_\alpha} \Phi_0 = \inf_{\mathcal{A}} \Phi$, i.e.

$$\begin{aligned}\mathcal{U}(\mathcal{A}_\alpha) &= \mathcal{U}(\mathcal{A}),\\ \mathcal{L}(\mathcal{A}_\alpha) &= \mathcal{L}(\mathcal{A}).\end{aligned}$$

For a subset $\Pi_0 \subseteq \mathcal{M}(\mathcal{A}_0)$, the projection identity (6.9) implies that the set $\Pi := P_0^{-1}\Pi_0$ defined by $P_0^{-1}\Pi_0 := \{\pi \in \mathcal{M}(\mathcal{A}) \mid P_0\pi \in \Pi_0\}$ is the induced set of probability measures on $\mathcal{A}$. Moreover, for $\pi \in \Pi$, the change of variables formula

$$\mathbb{E}_\pi[\Phi] = \mathbb{E}_\pi[\Phi_0 \circ P_\alpha] = \mathbb{E}_{P_\alpha\pi}[\Phi_0]$$

implies that

$$\sup_{\pi\in\Pi}\mathbb{E}_\pi[\Phi] = \sup_{\pi_\alpha\in P_\alpha\Pi}\mathbb{E}_{\pi_\alpha}[\Phi_0],$$
$$\inf_{\pi\in\Pi}\mathbb{E}_\pi[\Phi] = \inf_{\pi_\alpha\in P_\alpha\Pi}\mathbb{E}_{\pi}[\Phi_0],$$

so that

$$P_\alpha\Pi = P_\alpha P_0^{-1}\Pi_0 \subseteq \mathcal{M}(\mathcal{A}_\alpha)$$

is the induced set of probability measures on $\mathcal{A}_\alpha$. Let us denote this induced set by

$$\Pi_\alpha := P_\alpha P_0^{-1}\Pi_0 \tag{6.13}$$

so that these equalities become

$$\mathcal{U}(\Pi) = \mathcal{U}(\Pi_\alpha),$$
$$\mathcal{L}(\Pi) = \mathcal{L}(\Pi_\alpha).$$

For conditioning on observations, define $\mathbb{D}_0^n$ as in (6.5) and pull it back to the data map $\mathbb{D}^n\colon \mathcal{M}(\mathcal{X})\times\mathcal{M}(\mathcal{X})\to\mathcal{M}(\mathcal{X}^n)$ defined by $\mathbb{D}^n := \mathbb{D}_0^n\circ P_\alpha$. Define B_δ^n as in (6.6) and recall the definition (5.2)

$$\mathbb{E}_{\pi\odot\mathbb{D}^n}\left[\Phi\middle|B_\delta^n\right] = \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\Phi(\mu_1,\mu_2)\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right]}.$$

of the conditional expectation and the corresponding (5.6) upper value

$$\mathcal{U}(\Pi\odot_{B_\delta^n}\mathbb{D}^n) := \sup_{\pi\odot\mathbb{D}^n\in\Pi\odot_{B_\delta^n}\mathbb{D}^n}\mathbb{E}_{\pi\odot\mathbb{D}^n}\left[\Phi\middle|B_\delta^n\right]$$

in terms of the admissible set (5.4)

$$\Pi\odot_{B_\delta^n}\mathbb{D}^n := \left\{\pi\odot\mathbb{D}^n : \pi\in\Pi,\ (\pi\cdot\mathbb{D}^n)[B_\delta^n] > 0\right\}$$

of product measures, where the marginal is defined by

$$(\pi\cdot\mathbb{D}^n)[B_\delta^n] := \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\left[\mathbb{D}^n(\mu_1,\mu_2)[B_\delta^n]\right].$$

Let us indicate the dependence on some measure $\underline{\pi}$ of the essential supremum of some quantity of interest $\underline{\Phi}$ by

$$\underline{\pi}^\infty(\underline{\Phi}) := \inf\{r\in\mathbb{R}\,|\,\underline{\pi}\{\underline{\Phi} > r\} = 0\}$$

and for a set $\underline{\Pi}$ of measures

$$\underline{\Pi}^\infty(\underline{\Phi}) := \sup_{\underline{\pi}\in\underline{\Pi}}\underline{\pi}^\infty(\underline{\Phi}). \tag{6.14}$$

For $\pi_\alpha = P_\alpha \pi$ with $\pi \in \Pi$, we have

$$\begin{aligned}\pi_\alpha[\Phi_0 > r] &= \big(P_\alpha \pi\big)[\Phi_0 > r] \\ &= \pi[\Phi_0 \circ P_\alpha > r] \\ &= \pi[\Phi > r]\end{aligned}$$

so that we conclude that

$$\Pi^\infty(\Phi) = \Pi^\infty_\alpha(\Phi_0).$$

Let us now quantify a type of regularity for the model $\mathcal{P}$. For $x \in \mathcal{X}$, let $B_0(x) := \{x\}$ and define

$$\mathcal{P}_\infty(\delta) := \sup_{x \in \mathcal{X}} \sup_{\theta \in \Theta} \mathcal{P}(\theta)[B_\delta(x)], \quad \text{for } \delta \geq 0.$$

It is clear that $\mathcal{P}_\infty \colon \mathbb{R}^+ \to [0,1]$ is an increasing function. Moreover, for most parametric families, it is easy to show that $\mathcal{P}_\infty$ is continuous and $\mathcal{P}_\infty(0) = 0$, and for many of them not difficult to find useful upper bounds.

Finally, let us assume that the model $\mathcal{P}$ is positive, in that $\mu(B_\delta(x)) > 0$ for all $\mu \in \mathcal{A}_0$, $x \in \mathcal{X}$, and $\delta > 0$.

Theorem 6.4 is a direct consequence of the following theorem.

Theorem 6.10 (Extreme Brittleness under Local Misspecification)**.** *With the notation and assumptions above, let Π_α be defined as in (6.13), and let $\delta > 0$ and $0 < \alpha < 1$ satisfy*

$$\mathcal{P}_\infty(\delta) < \alpha.$$

Then, for all integers $n \geq 1$,

$$\mathcal{U}(\Pi_\alpha \odot_{B^n_\delta} \mathbb{D}^n_0) \geq \Pi^\infty_0(\Phi_0)$$

with similar expressions for the lower bounds $\mathcal{L}$.

Remark 6.11. When Cromwell's rule (see Section 2) is implemented (i.e. if the prior measure of every non-empty neighborhood is strictly positive), it follows that $\Pi^\infty_0(\Phi_0) = \mathcal{U}(\mathcal{A}_0)$ so that the conclusion of Theorem 6.10 becomes

$$\mathcal{U}(\Pi_\alpha \odot_{B^n_\delta} \mathbb{D}^n_0) \geq \mathcal{U}(\mathcal{A}_0)\,.$$

Remark 6.12. Theorem 6.10 provides conditions sufficient to guarantee how bad things can get regardless of how many samples are taken. One might hope that when these conditions are not satisfied, that more samples may prove beneficial. However, when the condition

$$\inf_{(\mu,\mu') \in \Psi^{-1}\mu} \mathbb{D}^n(\mu,\mu')[B^n_\delta] = 0, \quad \mu \in \mathcal{A}_0$$

of Theorem 6.1 is only approximately satisfied, the inequality

$$\mathbb{D}^n(\mu,\mu')[B^n_\delta] = \big(\mu'\big)^n[B^n_\delta] = \prod_{i=1}^n \mu'[B_\delta(x_i)]$$

and the quantitative version of Theorem 5.12 (given in [85, Thm. 3.1], see also [85, Rmk. 3.2]) imply that things actually get worse with more samples.

Remark 6.13. The proof of Theorem 6.10 shows that we can obtain a similar result when using the total variation distance d_{TV} instead of the Prokhorov distance, which produces a much smaller neighborhood. However, in this metric $\mathcal{M}(\mathcal{X})$ in general is not separable and this introduces measurability difficulties. These difficulties can be overcome somewhat when Θ is compact and $\mathcal{P}$ is continuous, since the image of a compact set under a continuous map is compact and therefore measurable. Moreover, validation or certification type quantities of interest defined by $\Phi(\mu) := \mu(A)$ for some fixed $A \in \mathcal{B}(\mathcal{X})$ are easily seen to be continuous and therefore measurable. Moreover, because of continuity,

$$\Pi_0^\infty(\Phi_0) \approx \Pi_\alpha^\infty(\Phi_0).$$

Our motivation in working mainly with the Prokhorov metric lies in the fact that we also seek to lay down measurability foundations for the scientific computation of optimal statistical estimators where the unknown quantities are products of functions and measures and for such spaces the total variation metric is too strong for the measurability of standard quantities of interest.

7 Admissible Sets as Measurable Spaces

In the Kolmogorov formulation of probability, to put probability measures on an admissible set $\mathcal{A} \subseteq \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ requires that the admissible set be a measurable space, i.e. $\mathcal{A}$ must be equipped with a σ-algebra of subsets upon which measures can be defined. This section concerns the development of such measurable structures. We will first describe a simple measurable structure corresponding to a non-separable complete topological space. However, the non-separability appears to make much of our analysis difficult, and so we also develop measurable structures which come from Polish (completely metrizable separable) spaces, which appears to give us what we need- not only the ability to apply the reduction theorems of Owhadi et al. [86], but appears to satisfy the technical needs of developing the Bayesian OUQ framework. See also the discussion of the benefits of Polish spaces in Remark 3.2.

To begin, let $\mathcal{X}$ be a metrizable topological space, let $\mathcal{F}(\mathcal{X})$ be a set of real-valued functions on $\mathcal{X}$, and consider the space $\mathcal{M}(\mathcal{X})$ of Borel probability measures on $\mathcal{X}$ equipped with the weak-$*$ topology and the corresponding Borel σ-algebra $\mathcal{B}\big(\mathcal{M}(\mathcal{X})\big)$. To put a probability measure on a subset

$$\mathcal{A} \subseteq \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$$

requires defining a σ-algebra of subsets of $\mathcal{A}$. To do this in a way that is robust to the selection of the particular subset $\mathcal{A}$ it is sufficient and, depending on the nature of the set of permissible assumption sets, possibly necessary to specify a σ-algebra on $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ and induce a σ-algebra to a specified subset $\mathcal{A} \subset \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ through relativization. Let us first consider generalized moment constraints. By [4, Thm. 15.13], for a separable and metrizable space $\mathcal{X}$, and for any bounded measurable function $g \colon \mathcal{X} \to \mathbb{R}$, it follows that the map $\mathcal{M}(\mathcal{X}) \to \mathbb{R}$ defined by $\mu \mapsto \int g \, \mathrm{d}\mu$ is measurable. Therefore, arbitrary bounded generalized moment constraints are in general measurable.

However, the objective function for many OUQ problems is not quite so simple. Often it is of the form

$$\Phi(f,\mu) = \big(f_*\mu\big)(A), \quad \text{for } (f,\mu) \in \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$$

for some fixed measurable subset $A \subseteq \mathbb{R}$. Therefore, we can deduce a σ-algebra on $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ by pulling back the σ-algebra $\mathcal{B}(\mathbb{R})$ to $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ using the map $(f,\mu) \mapsto \big(f_*\mu\big)(A)$ and also pulling back under each of the maps $(f,\mu) \mapsto \mathbb{E}_\mu[g]$ for $g \in \mathcal{G}$ where $\mathcal{G}$ is the constraint set. However useful such a method might be for a particular problem, it is appealing to describe measurable structures on $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ for which a large class of objective functions Φ, in particular $\Phi_A(f,\mu) = \big(f_*\mu\big)(A)$ for various A, and constraints, would be generally measurable.

To that end, we first demonstrate that by restricting $\mathcal{F}(\mathcal{X})$ to the bounded measurable functions $B(\mathcal{X})$ equipped with the supremum norm, that the product structure for $B(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ makes objective functions Φ_A, $A \in \mathcal{B}(\mathbb{R})$, measurable and generalized moment constraints measurable. However, although the topological space $B(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is complete and metrizable, it is, in general, not separable. We then describe a general procedure for choosing smaller, but still very large sets of functions $\mathcal{F}(\mathcal{X})$ in such a way that $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is Polish and such that all objective functions $\Phi_A(f,\mu) = \big(f_*\mu\big)(A)$ for $A \in \mathcal{B}(\mathbb{R})$ and all generalized moment constraints are measurable. To that end, let us describe some notation. For a topological space $\mathcal{Z}$, let $B(\mathcal{Z})$ denote the Banach space of bounded measurable real valued functions on $\mathcal{Z}$ with the uniform (supremum) norm $\|\cdot\|_\infty$ and let $\mathcal{B}\big(B(\mathcal{Z})\big)$ denote the σ-algebra of subsets of $B(\mathcal{Z})$ corresponding to the metric topology of $B(\mathcal{Z})$. When $\mathcal{Z}$ is metric, let $\mathrm{BL}(\mathcal{Z})$ denote the space of bounded, Lipschitz, measurable real valued functions on $\mathcal{Z}$ with norm $\|\cdot\|_{\mathrm{BL}} := \|\cdot\|_\infty + \|\cdot\|_{\mathrm{Lip}}$.

For a set $\mathcal{F}(\mathcal{X})$ of real valued measurable functions on $\mathcal{X}$ equipped with a topology $\tau(\mathcal{F})$ we consider the map

$$J\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R}) \tag{7.1}$$

defined by

$$J(f,\mu) := f_*\mu, \quad \text{for } (f,\mu) \in \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}). \tag{7.2}$$

We will be interested in developing conditions on the topological space $\big(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F})\big)$ which guarantee that J is measurable, that is

$$J\colon \Big(\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}), \mathcal{B}\big(\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})\big)\Big) \to \Big(\mathcal{M}(\mathbb{R}), \mathcal{B}\big(\mathcal{M}(\mathbb{R})\big)\Big). \tag{7.3}$$

The goal of this section will be the proof of the following theorem:

Theorem 7.1. *Suppose that $\mathcal{X}$ is Polish and $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathbb{R})$ are equipped with the weak-$*$ topologies. When $\mathcal{X}$ is compact, consider $\mathcal{F}(\mathcal{X}) := C(\mathcal{X})$, the Banach space of continuous functions. Or more generally, let $\mathcal{F}(\mathcal{X})$ be a RKHS (Reproducing Kernel Hilbert Space) or RKBS (Reproducing Kernel Banach Space) of real functions with a measurable feature map, or the space $\mathrm{UC}(\mathcal{X})$ of upper semicontinuous functions with the Wijsman topology obtained through the identification of an upper semicontinuous function with its hypograph. Then $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is Polish and $J\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ defined by $J(f,\mu) := f_*\mu$ is measurable.*

7.1 Evaluation Measurable Function Spaces

In anticipation of proving measurability by proving that J is a Carathéodory function (one that is continuous in one variable and measurable in the other [4, Def. 4.50]), the following lemma is useful and should have independent interest: namely [4, Thm. 15.14] states that if $f\colon \mathcal{X} \to \mathcal{Y}$ is continuous, then $f_*\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ is continuous. We show, using a result of Kechris [71, Thm. 13.11, p. 84], that the continuity requirement of f can be removed while still obtaining continuity of f_*. Continuity with respect to μ is a large step towards the measurability of J: now all that remains is the measurability with respect to f.

Lemma 7.2. *Let $(\mathcal{X}, \tau_{\mathcal{X}})$ be Polish and $(\mathcal{Y}, \tau_{\mathcal{Y}})$ metrizable and second countable, and let $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathcal{Y})$ denote the corresponding spaces of Borel probability measures endowed with the weak-$*$ topologies. Let $f\colon (\mathcal{X}, \mathcal{B}(\tau_{\mathcal{X}})) \to (\mathcal{Y}, \mathcal{B}(\tau_{\mathcal{Y}}))$ be measurable. Then $f_*\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{Y})$ is continuous.*

We are now in a position to state our first measurability result.

Proposition 7.3. *Let $\mathcal{X}$ be Polish and consider $\mathcal{M}(\mathcal{X})$ and $\mathcal{M}(\mathbb{R})$ endowed with the weak-$*$ topologies. Consider $\mathcal{F}(\mathcal{X}) := B(\mathcal{X})$ endowed with the metric topology corresponding to $\|\cdot\|_\infty$. Then $B(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is complete metrizable and the map J, defined in (7.1) and (7.2), is measurable.*

Proposition 7.3 shows that J is measurable with respect to the Borel structure of the product space $B(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$. However, although the product space $B(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ is complete and metrizable, it is in general not separable. Indeed, it appears that this space is so large that separate continuity of $J\colon B(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ is available, suggesting that it might be possible to weaken this measurable structure by judicious choice of topological function space $\big(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F})\big)$ in such a way that makes $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ into a Polish space and keeps J measurable.

In Arens [5] a topology on a space of functions is called *admissible* if point evaluation is jointly continuous in the product of the space of functions and the domain. To achieve the aforementioned goal of making $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ Polish and keeping $J\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ measurable for a set $\mathcal{F}$ of measurable functions, we generalize Arens' notion from topological spaces to measurable spaces by requiring the identity map to be a *normal integrand* in the sense of Rockafellar and Wets [90, Def. 14.27]. Specifically,

Definition 7.4. A set $\mathcal{F}(\mathcal{X})$ of real valued measurable functions on a topological space $\mathcal{X}$ equipped with a σ-algebra $\sigma(\mathcal{F})$ of subsets of $\mathcal{F}(\mathcal{X})$ is called an *evaluation measurable function space* if the mapping $i_x\colon \big(\mathcal{F}(\mathcal{X}), \sigma(\mathcal{F})\big) \to \big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ defined by

$$i_x f := f(x), \quad \text{for } f \in \mathcal{F}(\mathcal{X})$$

is measurable for all $x \in \mathcal{X}$.

In many situations, it appears to be a minimal requirement that a measurable space $\big(\mathcal{F}(\mathcal{X}), \sigma(\mathcal{F})\big)$ be evaluation measurable. The following result shows that this is equivalent to the measurability of J in the product structure:

Theorem 7.5. *Suppose that $\mathcal{X}$ is Polish. Then $J\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ is measurable in the product structure $\sigma\big(\mathcal{F}\big) \times \mathcal{B}\big(\mathcal{M}(\mathcal{X})\big)$ if and only if $\big(\mathcal{F}(\mathcal{X}), \sigma(\mathcal{F})\big)$ is evaluation measurable.*

In particular, since point evaluation on $B(X)$ is continuous, it is measurable, so that Theorem 7.5 implies Proposition 7.3.

The following corollary says that composing the response function f with a utility function h maintains measurability:

Corollary 7.6. *Given the assumptions of Theorem 7.5, suppose that $h\colon \mathbb{R} \to \mathbb{R}$ is measurable. Then the map $J_h\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ defined by*

$$J_h(f, \mu) := (h \circ f)_* \mu, \quad \text{for } (f, \mu) \in \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$$

is measurable in the product structure.

Furthermore, when $\big(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F})\big)$ is Polish, Theorem 7.5 and Corollary 7.6 provide measurability on the the product space as follows:

Corollary 7.7. *Suppose that $\mathcal{X}$ and $\big(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F})\big)$ are Polish. Then $J\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ is measurable in the Borel structure $\mathcal{B}\big(\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})\big)$ of the product space if and only if $\big(\mathcal{F}(\mathcal{X}), \mathcal{B}(\tau(\mathcal{F}))\big)$ is evaluation measurable.*

Moreover, if $\big(\mathcal{F}(\mathcal{X}), \mathcal{B}(\tau(\mathcal{F}))\big)$ is evaluation measurable, then for any measurable function $h\colon \mathbb{R} \to \mathbb{R}$, the map $J_h\colon \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ defined by

$$J_h(f, \mu) := (h \circ f)_* \mu, \quad \text{for } (f, \mu) \in \mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$$

is measurable in the Borel structure of the product space.

7.2 Polish Evaluation Measurable Function Spaces

As a consequence of Corollary 7.7, it follows that to make $\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ a Polish space such that J is measurable, it is necessary and sufficient that $\mathcal{F}(\mathcal{X})$ be a Polish evaluation measurable function space with respect to its Borel structure. Fortunately, such spaces already have been well studied in the literature.

The Banach space $C(\mathcal{X})$ of bounded continuous functions is, by [4, Lem. 3.99], separable when $\mathcal{X}$ is compact and metrizable. Since point evaluation is continuous in general, when $\mathcal{X}$ is compact metrizable $C(\mathcal{X})$ is then a Polish evaluation measurable function space. When $\mathcal{X}$ is not compact, this is not the case.

Reproducing Kernel Hilbert Spaces (RKHS), see e.g. [101, Sec. 4] and [23], are extremely important in Learning Theory. Unlike the Lebesgue space L^2, a RKHS H is a Hilbert space of real valued functions — not equivalence classes — characterized by the fact that point evaluation is a continuous function on the Hilbert space. Consequently, any separable RKHS of functions on $\mathcal{X}$ is a Polish evaluation measurable function space. To obtain separability, [101, Lem. 4.33] asserts that if $\mathcal{X}$ is separable and the kernel k corresponding to the RKHS H is continuous, then H is separable. More generally,

Steinwart and Scovel [105, Cor. 3.6] show that if there exists a finite and strictly positive Borel measure on $\mathcal{X}$, then every bounded and separately continuous kernel k has a separable RKHS H. Also, [23, Thm. 15, pg. 33] shows that RKHS H is separable if there is a countable subset $\mathcal{X}_0 \subseteq \mathcal{X}$ such that if $f \in H$ and $f(x) = 0$ for all $x \in \mathcal{X}_0$ then $f = 0$. Finally, a result of Fortet [52, Thm. 1.2] asserts that a RKHS H with kernel k is separable if and only if for all $\varepsilon > 0$ there exists a countable partition $B_j, j \in \mathbb{N}$ of $\mathcal{X}$ such that for all $j \in \mathbb{N}$ and all $x_1, x_2 \in B_j$ we have

$$k(x_1, x_1) + k(x_2, x_2) - k(x_1, x_2) - k(x_2, x_1) < \varepsilon.$$

A RKHS is usually implicitly defined by its kernel and often it is desirable to have a more concrete representation of the corresponding RKHS H. Although important RKHSs such as the Fock space described by Bargmann [9] have been known for some time, it is only recently that Steinwart, Hush and Scovel [102] provided an explicit description of Gaussian RKHSs. However, even without an explicit representation, often we can say something about how expressive the RKHS is in terms of approximation properties. Steinwart [100] introduced *universal kernels* on compact domains as those whose RKHS can approximate any continuous function uniformly, and demonstrated that many of the existing popular kernels, in particular the Gaussian RKHSs, are universal. For noncompact $\mathcal{X}$, Steinwart, Hush and Scovel [103] provide conditions on the kernel which guarantee approximation properties in L^p spaces. Most important however, is that the expressive capability of the Gaussian RKHSs is part of what allowed Steinwart and Scovel [104] to prove that support vector machines learn fast. For a thorough discussion of these topics, see [101]. Although the current investigated approximation properties are with respect to Learning Theory, they suggest that approximation properties with respect to, for example, $J(f, \mu) = f_*\mu$ might be available.

Reproducing Kernel Banach Spaces (RKBS), introduced by Zhang, Xu, and Zhang [122] are Banach spaces of real valued functions for which point evaluation is continuous. Therefore, any separable Reproducing Kernel Banach Space is a linear Polish evaluation measurable function space. An "if and only if" characterization of separability is obtained through a generalization of Fortet's Theorem from RKHSs to RKBSs. We suspect our proof is similar to Fortet's for RKHSs, but it is not written down in [52]. Indeed, Fortet's result mentioned above, is a regularity condition on the pullback metric

$$d_H(x_1, x_2) := \|\Phi(x_1) - \Phi(x_2)\|_{H_1} = \sqrt{k(x_1, x_1) + k(x_2, x_2) - k(x_1, x_2) - k(x_2, x_1)}$$

to $\mathcal{X}$ determined by a feature map $\Phi\colon \mathcal{X} \to H_1$. In particular, Fortet's condition then becomes: for all $j \in \mathbb{N}$ and all $x_1, x_2 \in B_j$ we have

$$d_H(x_1, x_2) < \sqrt{\varepsilon}.$$

We refer to [122] for the foundational facts and terminology regarding RKBSs.

Lemma 7.8 (Fortet)**.** *A RKBS B is separable if and only if there exists a feature Banach space $\mathcal{W}$ and feature map $\Phi\colon \mathcal{X} \to \mathcal{W}$ for B such that for all $\varepsilon > 0$ there is a countable*

partition $B_j \subset \mathcal{X}, j \in \mathbb{N}$ *with* $\bigcup_{j\in\mathbb{N}} B_j = \mathcal{X}$ *such that for all* $j \in \mathbb{N}$ *and all* $x_1, x_2 \in B_j$, *we have*

$$\|\Phi(x_1) - \Phi(x_2)\|_{\mathcal{W}} < \varepsilon. \tag{7.4}$$

From Lemma 7.8, it follows that if $\mathcal{X}$ is Lindelöf (meaning that every open cover has a countable subcover), then any RKBS of functions on $\mathcal{X}$ which has a continuous feature map $\Phi\colon \mathcal{X} \to \mathcal{W}$ is separable. Therefore, since Polish implies Suslin implies Lindelöf, RKBSs of functions on Polish or Suslin spaces are separable when there is a continuous feature map. Moreover, from the proof of Lemma 7.8, we easily conclude the RKBS version of [101, Lem. 4.33] when combined with [101, Lem. 4.29]: a RKBS of functions on a separable space $\mathcal{X}$ is separable if it has a continuous feature map.

Finally, a very strong characterization of separability is available, due to a theorem of Stone [106, Thm. 16, pg. 32], when $\mathcal{X}$ is a separable absolutely Borel space, in particular, when $\mathcal{X}$ is Polish. Following Frolik [57], a metrizable space $\mathcal{X}$ is said to be *absolutely Borel* if $\mathcal{X} \subset \mathcal{Z}$ is a Borel subset for all metrizable $\mathcal{Z}$. Moreover, Frolik [56] introduces *bianalytic spaces* as analytic spaces such that their complement in their Čech compactification is also analytic and, in Frolik [56, Thm. 12], shows that a metrizable space is separable absolute Borel if and only if it is bianalytic. The following result regarding the separability of RKHS and RKBS is of independent interest and easily gives us our main result when $\mathcal{X}$ is Polish.

Lemma 7.9. *Let* $\mathcal{X}$ *be bianalytic and let* $\mathcal{K}$ *be a RKHS with measurable feature map or a RKBS with measurable primary feature map. Then* $\mathcal{K}$ *is separable.*

Theorem 7.10. *Let* $\mathcal{X}$ *be Polish and let* $\mathcal{K}$ *be a RKHS with measurable feature map or a RKBS with a dual pair of measurable feature maps. Then* $\mathcal{K}$ *is a Polish evaluation measurable function space.*

The space $D(\mathcal{X})$ of differences of upper semicontinuous functions has been investigated by Rosenthal [91] who describes a Banach space structure for it. However, regarding separability, Rosenthal [92] tells us that "off the top, I'd say — almost never. e.g. if $\mathcal{X}$ has a non-trivial convergent sequence".

The space $\mathrm{DC}(\mathcal{X})$ of differences of convex functions, see Tuy [109], is important in non-convex optimization because if the decomposition into the difference of convex functions is known then convex analysis can be used for both the algorithms and the development and evaluation of optimality criteria. Moreover, the class $\mathrm{DC}(\mathcal{X})$ is relatively rich. For example, Tuy [109, Prop. 3.2] states that the restriction f_Ω of any twice continuously differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ to a compact convex set $\Omega \subset \mathbb{R}^n$ is in $\mathrm{DC}(\Omega)$. At present, we are not aware of any topologies for either $D(\mathcal{X})$ or $\mathrm{DC}(\mathcal{X})$ that would make them Polish.

The space $\mathrm{UC}(\mathcal{X})$ of upper semicontinuous functions is not linear, but a cone. Semicontinuous functions are important in many areas of mathematics, in particular, optimization theory, and since the OUQ framework is optimization based, it appears natural to consider it. In the following sections we will describe a topology for the space $\mathrm{UC}(\mathcal{X})$ which makes it into a Polish evaluation measurable function space. Note that, unlike the

above examples, here point evaluation will *not* be continuous, but semicontinuous — and therefore still measurable!. We conjecture that this topology can be used to topologize $D(\mathcal{X})$ and $\mathrm{DC}(\mathcal{X})$ in such a way as to make them Polish evaluation measurable function spaces, but we leave that for the future.

7.3 Polish Topologies for Upper Semicontinuous Functions

Following Beer [12], we topologize the space of upper semicontinuous functions $\mathrm{UC}(\mathcal{X})$ through the identification of an upper semicontinuous function with its hypograph, which is a closed set, and a topology on the space of closed sets. To that end, recall that an upper semicontinuous function $f\colon \mathcal{X} \to \mathbb{R}$ is such that its hypograph

$$\mathrm{hypo}(f) := \{(x, \alpha) \in \mathcal{X} \times \mathbb{R} \mid f(x) \geq \alpha\}$$

is a closed set. An equivalent definition is that the excursion set

$$\{x \in \mathcal{X} \mid f(x) \geq \alpha\}$$

is closed for all $\alpha \in \mathbb{R}$. It follows then that upper semicontinuous functions are measurable. The mapping $f \mapsto \mathrm{hypo}(f)$ can be used to pull back structures from the set of closed convex sets to sets of upper semicontinuous functions. Let us denote the space of upper semicontinuous functions on $\mathcal{X}$ by $\mathrm{UC}(\mathcal{X})$ and the space of closed subsets of $\mathcal{X} \times \mathbb{R}$ by $\mathrm{CL}(\mathcal{X} \times \mathbb{R})$. Then we have a map

$$\mathrm{hypo}\colon \mathrm{UC}(\mathcal{X}) \to \mathrm{CL}(\mathcal{X} \times \mathbb{R})$$

and so if we topologize, metrize, or measurablize $\mathrm{CL}(\mathcal{X} \times \mathbb{R})$ we can pull back such structures to $\mathrm{UC}(\mathcal{X})$ through the map hypo.

7.4 Hyperspace Topologies and Measurability

Since the map $\mathrm{hypo}\colon \mathrm{UC}(\mathcal{X}) \to \mathrm{CL}(\mathcal{X}\times\mathbb{R})$ gives a method for transferring structure from spaces of closed subsets, we now describe topologies and σ-algebras for the hyperspace of closed subsets. This subject has been heavily researched and we will only define and use what we need. Evidently, the classic reference is Matheron [80]. Lest one should think that these ideas come from the ivory tower, note that Matheron developed these ideas for mining, see Agterberg [3] for an illuminating biography. For each of the increasing categories-compact, locally compact, Polish, for the base space $\mathcal{X}$ we will establish topologies on the the space $\mathrm{UC}(\mathcal{X})$ of upper semicontinuous functions making it a Polish evaluation measurable function space. Moreover, we will do so in one stroke. For a topological space $\mathcal{X}$ let $\mathcal{G}$ denote the collection of non-empty open sets, $\mathcal{F}$ the collection of non-empty closed sets, $\mathcal{K}$ the collection of non-empty compact sets.

The most famous hyperspace topology is the Vietoris topology [82] but we will not use it, except to note that by [4, Thm. 3.91] that the Vietoris topology τ_V and the Hausdorff metric topology τ_h, to be defined below, coincide when relativized to $\mathcal{K}$. Let

us proceed to the Hausdorff metric topology. When $(\mathcal{X}, d)$ is a semimetric space we can define the distance from a point $x \in \mathcal{X}$ to a subset $A \subseteq \mathcal{X}$ by

$$d(x, A) := \inf_{x' \in A} d(x, x').$$

and the Hausdorff distance between arbitrary subsets by

$$h_d(A, B) := \max \left\{ \sup_{x \in A} d(x, B), \sup_{x' \in B} d(x', A) \right\}, \quad \text{for } A, B \subseteq \mathcal{X}. \tag{7.5}$$

By [4, Lem. 3.74] we have a characterization which, among other things, allows a direct comparison of the Hausdorff metric topology with the Wijsman topology which we will describe soon:

$$h_d(A, B) = \sup_{x \in \mathcal{X}} \big| d(x, A) - d(x, B) \big|, \quad \text{for } A, B \subseteq \mathcal{X}.$$

When d is a metric, it follows that h_d is an extended valued metric on the set of closed subsets $\mathcal{F}$ and as such by [4, Lem. 3.77] defines a first countable Hausdorff metric topology τ_h on $\mathcal{F}$. Moreover, by [4, Thm. 3.91] when relativized to $\mathcal{K}$ the Hausdorff metric topology is topological in that it is the same for all admissible metrics for $\mathcal{X}$ metrizable. For our purposes, we need the fact [4, Thm. 3.85] that for a metric space $(\mathcal{X}, d)$,

$$(\mathcal{F}, \tau_h) \text{ is Polish if and only if } (\mathcal{X}, d) \text{ is compact} \tag{7.6}$$

from which we conclude that

$$\mathcal{X} \text{ is compact metrizable } \implies (\mathcal{F}, \tau_h) \text{ is Polish} \tag{7.7}$$

and the same, for all admissible metrics d. Consequently, when a metrizable $\mathcal{X}$ is not compact, then the Hausdorff metric topology on $\mathcal{F}$ defined by any admissible metric is not Polish.

When $\mathcal{X}$ is metrizable but not compact, let us consider instead the Fell topology [51]. It is defined as the topology τ_F generated by the base consisting of

$$\{F \in \mathcal{F} \mid F \cap G \neq \varnothing\}, \qquad G \in \mathcal{G} \tag{7.8}$$

$$\{F \in \mathcal{F} \mid F \cap K = \varnothing\}, \qquad K \in \mathcal{K}. \tag{7.9}$$

For our purposes, we need the fact [4, Cor. 3.95] that for a locally compact Polish space $\mathcal{X}$, we have

$$\mathcal{X} \text{ locally compact and Polish } \implies (\mathcal{F}, \tau_F) \text{ is Polish.} \tag{7.10}$$

For the other direction, Molchanov [83, Thm. B.2.iii] asserts that when $\mathcal{X}$ is Hausdorff

$$(\mathcal{F}, \tau_F) \text{ is Polish } \implies \mathcal{X} \text{ is locally compact and second countable.} \tag{7.11}$$

Moreover, [4, Thm. 3.93] implies that

$$\mathcal{X} \text{ is compact and metrizable } \implies \tau_h = \tau_F \tag{7.12}$$

for any admissible metric d. Therefore, the Fell topology can be considered a Polish generalization of the Hausdorff metric topology from compact metrizable $\mathcal{X}$ to locally compact Polish $\mathcal{X}$. However, for Hausdorff spaces, (7.11) asserts that this Polish generalization does not go past locally compact second countable spaces.

To get past this to infinite dimensions, let us consider the Wijsman [120] topology τ_W on a metric space $(\mathcal{X}, d)$, defined as the initial topology generated by the functions

$$A \mapsto d(x, A), \quad \text{for } A \in \mathcal{F} \tag{7.13}$$

as x varies over $\mathcal{X}$. It is the weakest topology on $\mathcal{F}$ such that the function (7.13) on $\mathcal{F}$ is continuous for all $x \in \mathcal{X}$. Wijsman [120] demonstrated that this topology makes the Fenchel transform continuous on locally compact spaces. Moreover, even though his results were stated for convex functions and sets, it is clear that much of his results carry over to non-convex sets and functions. A complete investigation of this topology, and its history, can be found under the name now used, Γ-convergence, in Dal Maso [39], where in particular, Γ-convergence is shown to be a powerful tool in homogenization.

However, what we use here is the following result of Beer [15]:

$$\mathcal{X} \text{ is Polish} \implies (\mathcal{F}, \tau_W) \text{ is Polish.} \tag{7.14}$$

To show that the Wijsman topology is a Polish generalization of the Fell topology, observe that Beer [13, Thm. 2.3] shows that when $\mathcal{X}$ is metric, we have $\tau_W = \tau_F$ if and only if $\mathcal{X}$ has nice closed balls, i.e. the only non-compact closed ball is the whole space. Moreover, Beer [14, Thm. 2] shows that when $\mathcal{X}$ is metrizable, it is locally compact if and only if it admits a metric with nice closed balls. Consequently, when $\mathcal{X}$ is locally compact and metrizable we have $\tau_W = \tau_F$ for any admissible metric. On the other hand, Beer [13, Pg. 92] shows that if a metric space has nice closed balls, then it is complete. Consequently, since a metric space with nice closed balls is clearly locally compact, we have for metric $\mathcal{X}$ that $\tau_W = \tau_F$ implies that $\mathcal{X}$ is locally compact Polish and therefore we conclude

$$\mathcal{X} \text{ locally compact Polish} \iff \tau_W = \tau_F \tag{7.15}$$

for all admissible metrics.

For a thorough investigation into the interrelationships between the Vietoris, Hausdorff, Fell, Wijsman and other hyperspace topologies for various topological spaces $\mathcal{X}$, see Beer, Lechicki, Levi, and Naimpally [17]. In particular, note that they show [17, Thm. 3.1] that for a metrizable space $\mathcal{X}$ that the Vietoris topology is the supremum of the Wijsman topologies over all admissible metrics, that is, the weakest topology such that $A \mapsto d(x, A), A \in \mathcal{F}$ is continuous for all $x \in \mathcal{X}$ and all admissible metrics d. On the other hand, when $\mathcal{X}$ is locally compact, Beer [14] shows the infima of the Wijsman topologies over the admissible metrics is the Fell topology. More generally, Costantini, Levi, and Pelanta [36, Thm. 3.1] show that the infima of the Wijsman topologies is the topology of upper Kuratowski convergence.

7.4.1 The Effros σ-algebra

Now let us discuss measurability on $\mathcal{F}$. The Effros [49] σ-algebra $\sigma_E(\mathcal{F})$ on $\mathcal{F}$ is defined to be the σ-algebra generated by sets of the form

$$\{F \in \mathcal{F} \mid F \cap G \neq \varnothing \text{ for } G \in \mathcal{G}\}. \tag{7.16}$$

Beer [15, Pg. 1125] credits Hess [64, 65] with the result that

$$\mathcal{X} \text{ is separable metric} \implies \sigma_E(\mathcal{F}) = \mathcal{B}(\tau_W), \tag{7.17}$$

i.e. the Effros σ-algebra is generated by the Wijsman topology. We summarize the above discussion in the following proposition:

Proposition 7.11. *Consider three admissible cases:*

$$\mathcal{X} = \begin{cases} \textit{compact metrizable with Hausdorff metric topology on } \mathcal{F} \\ \textit{locally compact Polish with Fell topology on } \mathcal{F} \\ \textit{Polish with Wijsman topology on } \mathcal{F} \end{cases}$$

In all three cases, the same hyperspace topology results when we use the Wijsman topology. That is, the above is the same as

$$\mathcal{X} = \begin{cases} \textit{compact metrizable with Wijsman topology on } \mathcal{F} \\ \textit{locally compact Polish with Wijsman topology on } \mathcal{F} \\ \textit{Polish with Wijsman topology on } \mathcal{F}\,. \end{cases}$$

Moreover, in all three cases the hyperspace topologies are Polish and the Borel σ-algebra corresponding with these topologies is the Effros σ-algebra.

7.5 Main Theorem for Semicontinuous Functions

We are now ready to state our main theorem.

Theorem 7.12. *Suppose that $\mathcal{X}$ is Polish, and consider the hyperspace $\mathrm{CL}(\mathcal{X} \times \mathbb{R})$ of closed subsets with the Wijsman topology τ_W. Let $\mathrm{UC}(\mathcal{X})$ denote the upper semicontinuous functions and $\mathrm{hypo}\colon \mathrm{UC}(\mathcal{X}) \to \mathrm{CL}(\mathcal{X} \times \mathbb{R})$ be the hypograph mapping. Consider the pullback topology $\tau_W(UC) := \mathrm{hypo}^{-1}(\tau_W)$ and its resulting Borel σ-algebra $\mathcal{B}(\tau_W(UC))$ of subsets of $\mathrm{UC}(\mathcal{X})$. Then*

$$\Big(\mathrm{UC}(\mathcal{X}), \mathcal{B}\big(\tau_W(UC)\big)\Big)$$

is a Polish evaluation measurable function space.

Remark 7.13. For an upper semicontinuous function f, define ${}_\rho f\colon \mathcal{X} \to \mathbb{R}$ by

$$\big({}_\rho f\big)(x) := \sup_{y \in B_\rho(x)} f(y).$$

Wijsman's Theorem [120, Thm. 6.1] states that when $\mathcal{X}$ has nice closed balls, then $\operatorname{hypo}(f_n) \to \operatorname{hypo}(f)$ if and only if

$$f = \lim_{\rho \to 0} \limsup_{n \to \infty} {}_\rho f_n$$
$$f = \lim_{\rho \to 0} \liminf_{n \to \infty} {}_\rho f_n$$

Although it appears that [120, Thm. 6.1] may be correct for Polish spaces that are not locally compact, the proof utilizes [120, Thm. 3.1] which required that the space have nice closed balls. However, the existence of this extension to non locally compact spaces does not affect the assertion of Theorem 7.12.

Corollary 7.14. *In each of the three admissible cases for $\mathcal{X}$ and the specified topology τ on $\mathrm{UC}(\mathcal{X})$ defined in Proposition 7.11, it follows that $\big(\mathrm{UC}(\mathcal{X}), \mathcal{B}(\tau)\big)$ is a Polish evaluation measurable function space.*

8 Proofs

8.1 Proof of Theorem 4.6

For $q \in \mathbb{R}^n$, define

$$\Pi(q) := \Psi^{-1}\mathfrak{Q} = \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\Psi] = q\}$$

and let $\Pi(q, n) := \Pi(q) \cap \Delta(n) \subseteq \Pi(q)$ be the subset consisting of $(n+1)$-fold convex combinations of Dirac masses. Using a layercake approach, we use the fact that

$$\Pi(Z) = \bigcup_{q \in Z} \Pi(q) \quad \text{and} \quad \Pi(Z, n) = \bigcup_{q \in Z} \Pi(q, n),$$

while applying Theorem 4.4 with equality constraints $\Pi(q), q \in \mathbb{R}^n$, and the fact that the supremum over a union is a supremum of suprema to obtain a reduction as follows:

$$\begin{aligned}
\mathcal{U}\big(\Pi(Z)\big) &= \mathcal{U}\left(\bigcup_{q \in Z} \Pi(q)\right) \\
&= \sup_{q \in Z} \mathcal{U}\big(\Pi(q)\big) \\
&= \sup_{q \in Z} \mathcal{U}\big(\Pi(q, n)\big) \\
&= \mathcal{U}\left(\bigcup_{q \in Z} \Pi(q, n)\right) \\
&= \mathcal{U}\big(\Pi(Z, n)\big).
\end{aligned}$$

8.2 Proof of Lemma 4.10

Since $T \subset \mathcal{Q}$ is a subset of a separable metrizable space, [4, Cor. 3.5] implies that it is itself separable and metrizable. Consider the set-valued map with non-empty values $\Psi^{-1}\colon T \twoheadrightarrow \mathcal{A}$ with graph G defined by

$$G := \left\{ \big(q, (f,\mu)\big) \in T \times \mathcal{A} \,\middle|\, \Psi(f,\mu) = q \right\}. \tag{8.1}$$

Let d be a metric that generated the topology of T and define $h\colon T \times \mathcal{A} \to \mathbb{R}$ by $h\big(q,(f,\mu)\big) := d(\Psi(f,\mu), q)$. Then, since d is continuous in each of its arguments, it follows that h is a Carathéodory function, as defined in Definition 9.2. Since T is separable and metrizable, Lemma 9.3 implies that h is $\mathcal{B}(T) \otimes \mathcal{B}(\mathcal{A})$-measurable. Rewriting Equation (8.1) as

$$G := \left\{ \big(q, (f,\mu)\big) \in T \times \mathcal{A} \,\middle|\, h\big(q,(f,\mu)\big) = 0 \right\}$$

yields that G belongs to $\mathcal{B}(T) \otimes \mathcal{B}(\mathcal{A})$. Lemma 9.1 (through the identification $S = \mathcal{A}$, $s = (f,\mu)$, $\varphi(t,s) = \Phi(f,\mu)$) implies that the function $\mathcal{U} \circ \Psi^{-1}\colon T \to \mathbb{R}$ defined for $q \in T$ by $q \mapsto \sup_{(f,\mu)\in\Psi^{-1}(q)} \Phi(f,\mu)$ is $\hat{\mathcal{B}}(T)$-measurable, thereby establishing the first assertion. The second assertion then follows from the second part of Lemma 9.1.

8.3 Proof of Theorem 4.11

For the first assertion, consider $\mathbb{Q} \in \mathfrak{Q}$. Then, by the second assertion of Lemma 4.10, there exists a $\hat{\mathcal{B}}(\operatorname{supp}\mathbb{Q})$-measurable section ψ of Ψ, i.e. a $\hat{\mathcal{B}}(\operatorname{supp}(\mathbb{Q}))$-measurable function $\psi\colon \operatorname{supp}(\mathbb{Q}) \to \mathcal{A}$ such that $\Psi(\psi(q)) = q$ for all $q \in \operatorname{supp}(\mathbb{Q})$. Let $\mathbb{Q}$ also denote its restriction to its support and $\hat{\mathbb{Q}}$ its completion. Let $\pi := \psi\hat{\mathbb{Q}} \in \mathcal{M}(\mathcal{A})$, so that, for all $A \in \mathcal{B}(\operatorname{supp}(\mathbb{Q}))$,

$$\begin{aligned}
\big(\Psi\pi\big)(A) &= \big(\Psi\psi\hat{\mathbb{Q}}\big)(A) \\
&= \big((\Psi\circ\psi)\hat{\mathbb{Q}}\big)(A) \\
&= \hat{\mathbb{Q}}(A) \\
&= \mathbb{Q}(A).
\end{aligned}$$

Hence, $\Psi\pi = \mathbb{Q}$, establishing the first assertion.

For the main assertion, observe that, for all $(f,\mu) \in \mathcal{A}$,

$$\big(\mathcal{U} \circ \Psi^{-1} \circ \Psi\big)(f,\mu) = \sup_{(f',\mu')\colon \Psi(f',\mu') = \Psi(f,\mu)} \Phi(f',\mu') \geq \Phi(f,\mu). \tag{8.2}$$

Consequently, for $\mathbb{Q} \in \mathfrak{Q}$, the first assertion shows that there is a π such that $\Psi\pi = \mathbb{Q}$, so that a change of variables (Proposition 9.7) and the monotonicity properties (Proposition 9.4) of these integrals, together with the inequality (8.2), imply that

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] &= \mathbb{E}_{\Psi\pi}[\mathcal{U} \circ \Psi^{-1}] \\
&= \mathbb{E}_{\pi}[\mathcal{U} \circ \Psi^{-1} \circ \Psi] \\
&\geq \mathbb{E}_{\pi}[\Phi],
\end{aligned}$$

and therefore

$$\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \sup_{\pi \in \Psi^{-1}\mathbb{Q}} \mathbb{E}_{\pi}[\Phi].$$

Consequently,

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] = \mathcal{U}(\Psi^{-1}\mathfrak{Q})$$

and, in particular,

$$\sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] \geq \mathcal{U}(\Psi^{-1}\mathfrak{Q}). \tag{8.3}$$

To obtain the reverse inequality, for $\delta > 0$ consider $\mathbb{Q} \in \mathfrak{Q}$ and apply Lemma 4.10 to conclude that there exists a δ-optimal $\hat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable section of Ψ; that is, a $\hat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable function $\psi\colon \mathrm{supp}(\mathbb{Q}) \to \mathcal{A}$ such that $\Psi\big(\psi(q)\big) = q$ for all $q \in \mathrm{supp}(\mathbb{Q})$ and $\big(\Phi \circ \psi\big)(q) > \big(\mathcal{U} \circ \Psi^{-1}\big)(q) - \delta$ for all $q \in \mathrm{supp}(\mathbb{Q})$. Now let $\pi := \psi\hat{\mathbb{Q}} \in \mathcal{M}(\mathcal{A})$, and observe from the proof of the first assertion that $\Psi\pi = \mathbb{Q}$, and therefore $\pi \in \Psi^{-1}\mathbb{Q}$. Therefore, by a change of variables,

$$\begin{aligned}
\mathbb{E}_{\pi}[\Phi] &= \mathbb{E}_{\psi\hat{\mathbb{Q}}}[\Phi] \\
&= \mathbb{E}_{\hat{\mathbb{Q}}}[\Phi \circ \psi] \\
&> \mathbb{E}_{\hat{\mathbb{Q}}}[\mathcal{U} \circ \Psi^{-1}] - \delta.
\end{aligned}$$

Since, by definition, $\mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] := \mathbb{E}_{\hat{\mathbb{Q}}}[\mathcal{U} \circ \Psi^{-1}]$, it follows that

$$\begin{aligned}
\mathcal{U}(\Psi^{-1}\mathfrak{Q}) &= \sup_{\pi \in \Psi^{-1}\mathfrak{Q}} \mathbb{E}_{\pi}[\Phi] \\
&\geq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}] - \delta.
\end{aligned}$$

Since $\delta > 0$ was arbitrary, it follows that

$$\mathcal{U}(\Psi^{-1}\mathfrak{Q}) \geq \sup_{\mathbb{Q} \in \mathfrak{Q}} \mathbb{E}_{\mathbb{Q}}[\mathcal{U} \circ \Psi^{-1}].$$

Recalling the reverse inequality (8.3), we obain the main assertion.

The assertion of measure affinity follows from Lemma 9.9

For the assertion (5.11), define

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_{\pi}[\psi_i] = 0 \text{ for } i = 1, \ldots, n\}.$$

Let $\epsilon > 0$. Assume that $\sup_{\pi_+ \in \Pi_+} \mathbb{E}_{\pi_+}\big[\Phi\big] > \lambda$ and that $\pi_+ \in \Pi_+$ is such that $\mathbb{E}_{\pi_+}\big[\Phi\big] > \lambda$. Observe that $\pi := \pi_+/\pi_+(\mathcal{A})$ is an element of Π that satisfies $\mathbb{E}_{\pi}\big[\Phi - \lambda\psi_0\big] > 0$. Define Π_n as in (4.9) and apply [86, Thm. 4.1] to $\sup_{\pi \in \Pi} \mathbb{E}_{\pi}\big[\Phi - \lambda\psi_0\big]$ to conclude that there exists $\pi^* \in \Pi_n$ such that $\mathbb{E}_{\pi^*}\big[\Phi - \lambda\psi_0\big] > 0$. Since $\Phi - \lambda\psi_0 = (\varphi - \lambda)\psi_0$ and ψ_0 is positive, it also follows that $\mathbb{E}_{\pi^*}\big[\psi_0\big] > 0$. Writing $\pi^*_+ := \pi^*/\mathbb{E}_{\pi^*}[\psi_0]$ we obtain that $\pi^*_+ \in \Pi_{+,n}$ and $\mathbb{E}_{\pi^*_+}\big[\Phi\big] > \lambda$, which concludes the proof of (5.11).

8.4 Proof of Lemma 5.1

Consider the set

$$\mathcal{Y}' := \bigcup \{\mathcal{O}_y \,|\, \mathcal{O}_y \subseteq \mathcal{Y} \text{ is open and } \nu(\mathcal{O}_y) = 0\}\,.$$

First let us show that $E = \mathcal{Y}'$. To see this, first observe that trivially we have $E \subseteq \mathcal{Y}'$. Now suppose that $y \in \mathcal{Y}'$. Then there exists a $y' \in \mathcal{Y}$ and an open $\mathcal{O}_{y'} \ni y'$ such that $y \in \mathcal{O}_{y'}$ and $\nu(\mathcal{O}_{y'}) = 0$. Therefore, $y \in E$ and hence $E = \mathcal{Y}'$.

Now, since $\mathcal{Y}'$ is a union of open sets, it is open and therefore measurable. Moreover, since $\mathcal{Y}$ is strongly Lindelöf, it follows that $\mathcal{Y}'$ is Lindelöf and that the open cover of $\mathcal{Y}'$ by ν-null open sets used in the definition of $\mathcal{Y}'$ has a countable subcover, so that

$$\mathcal{Y}' = \bigcup_{i\in\mathbb{N}} \mathcal{O}_{y_i}$$

where each $\mathcal{O}_{y_i}$ is open and has $\nu(\mathcal{O}_{y_i}) = 0$. It follows that

$$\nu(E) = \nu(\mathcal{Y}') \leq \sum_{i\in\mathbb{N}} \nu(\mathcal{O}_{y_i}) = 0$$

and the proof is finished.

8.5 Proof of Theorem 5.7

The first assertion, (5.10), follows by layering the set of positive measures of finite total mass as $\bigcup_{r\in\mathbb{R}_+}\{r\mathcal{M}(\mathcal{A})\}$, using the fact that the supremum over a union is a supremum of suprema, and applying the reduction theorem [86, Thm. 4.1] in $r\mathcal{M}(\mathcal{A})$ separately.

For the second assertion, (5.11), define

$$\Pi := \{\pi \in \mathcal{M}(\mathcal{A}) \,|\, \mathbb{E}_\pi[\psi_i] = 0 \text{ for } i = 1, \ldots, n\}$$

Let $\epsilon > 0$. Assume that $\sup_{\pi_+\in\Pi_+} \mathbb{E}_{\pi_+}\big[\Phi\big] > \lambda$ and that $\pi_+ \in \Pi_+$ is such that $\mathbb{E}_{\pi_+}\big[\Phi\big] > \lambda$. Observe that $\pi := \pi_+/\pi_+(\mathcal{A})$ is an element of Π that satisfies $\mathbb{E}_\pi\big[\Phi - \lambda\psi_0\big] > 0$. Defining Π_n as in (4.9) and applying [86, Thm. 4.1] to $\sup_{\pi\in\Pi} \mathbb{E}_\pi\big[\Phi - \lambda\psi_0\big]$, we deduce that there exists $\pi^* \in \Pi_n$ such that $\mathbb{E}_{\pi^*}\big[\Phi - \lambda\psi_0\big] > 0$. Since $\Phi - \lambda\psi_0 = (\varphi - \lambda)\psi_0$ and ψ_0 is positive, it also follows that $\mathbb{E}_{\pi^*}\big[\psi_0\big] > 0$. Let $\pi^*_+ := \pi^*/\mathbb{E}_{\pi^*}[\psi_0]$ to obtain that $\pi^*_+ \in \Pi_{+,n}$ and $\mathbb{E}_{\pi^*_+}\big[\Phi\big] > \lambda$, which concludes the proof of (5.11).

8.6 Proof of Theorem 5.8

Observing that

$$\mathcal{U}(\Pi(q) \odot_B \mathfrak{D}) = \sup_{\mathbb{D}\in\mathfrak{D}} \mathcal{U}(\Pi(q) \odot_B \mathbb{D})$$

we may fix $\mathbb{D} \in \mathfrak{D}$. First, we prove that

$$\mathcal{U}(\Pi(q) \odot_B \mathbb{D}) = \sup_{\pi_+\in\Pi_+(q)} \mathbb{E}_{(f,\mu)\sim\pi_+}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big], \tag{8.4}$$

where $\Pi_+(q)$ is the set of positive finite measures π_+ on $\mathcal{A}$ such that $\mathbb{E}_{\pi_+}\big[\Psi(f,\mu)-q\big]=0$ and $\mathbb{E}_{\pi_+}\big[\mathbb{D}(f,\mu)[B]\big]=1$. To that end, first observe that

$$\mathcal{U}(\Pi(q)\odot_B\mathbb{D})=\sup_{\pi\in\Pi(q):\pi\odot\mathbb{D}[B]>0}\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]$$

and that, for any $\pi\in\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$,

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]=\frac{\mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]}. \tag{8.5}$$

Now consider $\pi\in\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$. Then $\pi_+:=\pi/\mathbb{E}_\pi\big[\mathbb{D}(f,\mu)[B]\big]$ is an element of $\Pi_+(q)$ and (8.5) implies that

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big]=\mathbb{E}_{(f,\mu)\sim\pi_+}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big].$$

Conversely, if $\pi_+\in\Pi_+(q)$, then $\pi:=\pi_+/\pi_+[\mathcal{A}]$ is an element of $\Pi(q)$ such that $\pi\odot\mathbb{D}[B]>0$ and

$$\mathbb{E}_{(f,\mu)\sim\pi_+}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]=\frac{\mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]}.$$

Since the above argument also shows that $\Pi(q)\odot_B\mathbb{D}$ is nonempty if and only if $\Pi_+(q)$ is nonempty, (8.4) follows. The right hand side of (8.4) is a linear program in π_+, so Theorem 5.7 implies that the supremum in π_+ can be achieved by assuming π_+ to be the weighted sum of at most $n+1$ Dirac masses, i.e. by assuming that

$$\pi_+=\sum_{i=0}^{n}\alpha_i\delta_{f_i,\mu_i} \tag{8.6}$$

This finishes the proof of Theorem 5.8.

8.7 Proof of Theorem 5.10

First let us show that, for $\lambda\in\mathbb{R}$, the statement that

$$\mathbb{E}_{\pi\odot\mathbb{D}}[\Phi|B]>\lambda,\quad \pi\odot\mathbb{D}\in\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D} \tag{8.7}$$

is equivalent to the statement that

$$\mathbb{E}_{(f,\mu)\sim\pi}\Big[\big(\Phi(f,\mu)-\lambda\big)\mathbb{D}(f,\mu)[B]\Big]>0. \tag{8.8}$$

To that end, assume (8.7) and observe that the definition (5.4) of $\Psi^{-1}(\mathfrak{Q})\odot_B\mathbb{D}$ implies that $\pi\cdot\mathbb{D}[B]>0$, where, by (5.5),

$$\pi\cdot\mathbb{D}[B]:=\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]. \tag{8.9}$$

Consequently, by (5.2),

$$\mathbb{E}_{\pi\odot\mathbb{D}}\big[\Phi\big|B\big] = \frac{\mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big]}{\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]} > \lambda,$$

and the denominator is strictly positive. Therefore,

$$\begin{aligned}&\mathbb{E}_{(f,\mu)\sim\pi}\big[\big(\Phi(f,\mu)-\lambda\big)\mathbb{D}(f,\mu)[B]\big]\\&\quad= \mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big] - \lambda\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]\\&\quad> 0,\end{aligned}$$

and (8.8) follows. Conversely, assume (8.8) and observe that $\pi\cdot\mathbb{D}[B] > 0$. To see this, observe that, if $\pi\cdot\mathbb{D}[B] = 0$, then (8.9) implies that $\mathbb{D}(f,\mu)[B] = 0$ π-a.s. and so

$$\mathbb{E}_{(f,\mu)\sim\pi}\big[\big(\Phi(f,\mu)-\lambda\big)\mathbb{D}(f,\mu)[B]\big] = 0,$$

which is a contradiction. Consequently, $\pi\cdot\mathbb{D}[B] > 0$ and dividing the assumption

$$\begin{aligned}&\mathbb{E}_{(f,\mu)\sim\pi}\big[\big(\Phi(f,\mu)-\lambda\big)\mathbb{D}(f,\mu)[B]\big]\\&\quad= \mathbb{E}_{(f,\mu)\sim\pi}\big[\Phi(f,\mu)\mathbb{D}(f,\mu)[B]\big] - \lambda\mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]\\&\quad> 0\end{aligned}$$

by $\pi\cdot\mathbb{D}[B] := \mathbb{E}_{(f,\mu)\sim\pi}\big[\mathbb{D}(f,\mu)[B]\big]$ throughout yields (8.7) and the equivalence is established.

Using this equivalence, the main assertion then follows from a direct application of Theorem 4.11. Finally, since Φ is semibounded, it follows that $(f,\mu)\mapsto\Phi(f,\mu)\mathbb{D}(f,\mu)[B]$ is semibounded and measurable, and the assertion of measure affinity follows from Lemma 9.9.

8.8 Proof of Theorem 5.12

Let us first establish that the assumptions of the theorem are well defined. To that end, note that Lemma 4.10 implies that $q\mapsto\inf_{(f,\mu)\in\Psi^{-1}(q)}\mathbb{D}(f,\mu)[B]$ is $\hat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable and hence (5.15) is well defined. Similarly (5.16) is well defined.

For the proof of the theorem, fix $\delta > 0$, let $\mathbb{Q}\in\mathfrak{Q}$ and $\mathbb{D}\in\mathfrak{D}$ satisfy the assumptions, and define $\lambda := \mathcal{U}(\mathcal{A}) - \delta$. Since $(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]$ is bounded and measurable, Lemma 4.10 implies that the function $q\mapsto\theta(q) := \sup_{(f,\mu)\in\Psi^{-1}(q)}(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]$ is $\hat{\mathcal{B}}(\mathrm{supp}(\mathbb{Q}))$-measurable. Moreover, (5.15) implies that the function θ is non-negative with $\mathbb{Q}$-probability one and (5.16) implies that θ is strictly positive on a subset of strictly positive $\mathbb{Q}$-measure. Hence,

$$E_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q)}(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]\right] = \mathbb{E}_{\mathbb{Q}}[\theta] > 0,$$

and, therefore,

$$\sup_{Q\in\mathfrak{Q},\mathbb{D}\in\mathfrak{D}} \mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q)} (\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]\right] > 0.$$

It then follows from Theorem 5.10 that $\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\odot_B\mathfrak{D}\big)\geq\lambda=\mathcal{U}(\mathcal{A})-\delta$. Since $\delta>0$ was arbitrary, it follows that $\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\odot_B\mathfrak{D}\big)\geq\mathcal{U}(\mathcal{A})$. Theorem 5.5 implies that

$$\mathcal{U}\big(\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}\big)\leq\mathcal{U}(\mathcal{A})$$

and the theorem follows.

8.9 Proof of Theorem 5.20

Defining a function of interest $\bar{\Phi}:=(\Phi-m)^2$, we observe that the assumptions on Φ of Theorem 5.20 imply that those of Theorem 5.12 are satisfied for $\bar{\Phi}$. Therefore, applying Theorem 5.12 to $\bar{\Phi}$ yields

$$\sup_{\pi\odot\mathbb{D}\in\Psi^{-1}(\mathfrak{Q})\odot_B\mathfrak{D}} \mathbb{E}_{\pi\odot\mathbb{D}}\left[(\Phi-m)^2\,\middle|\,B\right] = \sup_{(f,\mu)\in\mathcal{A}}\big(\Phi(f,\mu)-m\big)^2.$$

Since

$$\sup_{(f,\mu)\in\mathcal{A}}\big(\Phi(f,\mu)-m\big)^2 = \begin{cases}\big(\mathcal{U}(\mathcal{A})-m\big)^2 & \text{if } m\leq\frac{\mathcal{U}(\mathcal{A})+\mathcal{L}(\mathcal{A})}{2},\\ \big(m-\mathcal{L}(\mathcal{A})\big)^2 & \text{if } m\geq\frac{\mathcal{U}(\mathcal{A})+\mathcal{L}(\mathcal{A})}{2},\end{cases}$$

minimizing over m completes the proof.

8.10 Proof of Theorem 6.1

The proof follows from the proof of Theorem 5.12 as follows. Let $\delta>0$, and let $\mathbb{D}\in\mathfrak{D}$ and a measurable section ψ satisfy the assumptions. Define $\lambda:=\mathfrak{Q}^\infty(\Phi\circ\psi)-\delta$, and the universally measurable function $q\mapsto\theta(q):=\sup_{(f,\mu)\in\Psi^{-1}(q)}(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]$. Then assumption (6.1) implies that the function θ is non-negative. It follows from the definition (6.14) of $\mathfrak{Q}^\infty(\Phi\circ\psi)$, and $\lambda<\mathfrak{Q}^\infty(\Phi\circ\psi)$, that there is a $\mathbb{Q}\in\mathfrak{Q}$ such that $\Phi\circ\psi>\lambda$ with nonzero $\mathbb{Q}$-measure. Since

$$\begin{aligned}\theta(q) &= \sup_{(f,\mu)\in\Psi^{-1}(q)}(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]\\ &\geq (\Phi\circ\psi(q)-\lambda)\mathbb{D}(\psi(q))[B],\end{aligned}$$

the positivity assumption (6.2) implies that θ is positive on a subset of positive $\mathbb{Q}$-measure. Hence,

$$\mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q)}(\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]\right]=\mathbb{E}_{\mathbb{Q}}[\theta]>0$$

and, therefore,

$$\sup_{\mathbb{Q}\in\mathfrak{Q},\mathbb{D}\in\mathfrak{D}} \mathbb{E}_{q\sim\mathbb{Q}}\left[\sup_{(f,\mu)\in\Psi^{-1}(q)} (\Phi(f,\mu)-\lambda)\mathbb{D}(f,\mu)[B]\right] > 0.$$

It then follows from Theorem 5.10 that

$$\mathcal{U}\big(\Psi^{-1}\mathfrak{Q}\odot_B\mathfrak{D}\big) \geq \lambda = \mathfrak{Q}^{\infty}(\Phi\circ\psi)-\delta.$$

Since $\delta>0$ was arbitrary, the assertion is proved.

8.11 Proof of Theorem 6.10

We appeal to the corollary, Theorem 6.1, to Theorem 5.12. To that end, let $\mathcal{A}$ be defined as in (6.8), and let $\mathcal{Q}:=\mathcal{A}_0$, $\Psi:=P_0$, $\mathfrak{D}:=\{\mathbb{D}^n\}$.

Since $\mathbb{D}^n=\mathbb{D}^n_0\circ P_\alpha$ is a pull-back,

$$\begin{aligned}(\pi\cdot\mathbb{D}^n)[B^n_\delta] &= \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n(\mu_1,\mu_2)[B^n_\delta]\big]\\ &= \mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n_0\circ P_\alpha(\mu_1,\mu_2)[B^n_\delta]\big]\\ &= \mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\mathbb{D}^n_0(\mu_2)[B^n_\delta]\big]\\ &= (P_\alpha\pi\cdot\mathbb{D}^n_0)[B^n_\delta],\end{aligned}$$

from which we conclude that $(P_\alpha\pi\cdot\mathbb{D}^n_0)[B^n_\delta]>0$ if and only if $(\pi\cdot\mathbb{D}^n)[B^n_\delta]>0$, and so conclude

$$\Pi_\alpha\odot_{B^n_\delta}\mathbb{D}^n_0 = P_\alpha\big(\Pi\odot_{B^n_\delta}\mathbb{D}^n\big),$$

where P_α acts on each component in the natural way. Moreover since $\Phi=\Phi_0\circ P_\alpha$ is also a pull-back, for $\pi\in\Pi$, we have

$$\begin{aligned}\mathbb{E}_{\pi\odot\mathbb{D}^n}\big[\Phi\big|B^n_\delta\big] &= \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\Phi(\mu_1,\mu_2)\mathbb{D}^n(\mu_1,\mu_2)[B^n_\delta]\big]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n(\mu_1,\mu_2)[B^n_\delta]\big]}\\ &= \frac{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\Phi_0\circ P_\alpha(\mu_1,\mu_2)\cdot\mathbb{D}^n_0\circ P_\alpha(\mu_1,\mu_2)[B^n_\delta]\big]}{\mathbb{E}_{(\mu_1,\mu_2)\sim\pi}\big[\mathbb{D}^n_0\circ P_\alpha(\mu_1,\mu_2)[B^n_\delta]\big]}\\ &= \frac{\mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\Phi_0(\mu_2)\mathbb{D}^n_0(\mu_2)[B^n_\delta]\big]}{\mathbb{E}_{\mu_2\sim P_\alpha\pi}\big[\mathbb{D}^n_0(\mu_2)[B^n_\delta]\big]}\\ &= \mathbb{E}_{P_\alpha\pi\odot\mathbb{D}^n_0}\big[\Phi_0\big|B^n_\delta\big]\end{aligned}$$

and so we conclude that

$$\mathcal{U}(\Pi\odot_{B^n_\delta}\mathbb{D}^n) = \mathcal{U}(\Pi_\alpha\odot_{B^n_\delta}\mathbb{D}^n_0). \tag{8.10}$$

We will now need the following proposition

Proposition 8.1. *Consider $B\in\mathcal{B}(\mathcal{X})$. Then for $\mu\in\mathcal{M}(\mathcal{X})$ such that $\mu(B)<1$, we have*

$$d_{\mathrm{TV}}(\mu,\mu|_{B^c})\leq\mu(B).$$

Proof. For $A \in \mathcal{B}(\mathcal{X})$, we have

$$\begin{aligned}\mu(A) - \mu|_{B^c}(A) &= \mu(A) - \frac{\mu(A \cap B^c)}{\mu(B^c)} \\ &= \mu(A \cap B) + \mu(A \cap B^c) - \frac{\mu(A \cap B^c)}{\mu(B^c)} \\ &= \mu(A \cap B) - \frac{\mu(B)}{1 - \mu(B)} \mu(A \cap B^c)\end{aligned}$$

and therefore

$$\mu(A) - \mu|_{B^c}(A) \leq \mu(A \cap B) \leq \mu(B)$$

and

$$\begin{aligned}\mu(A) - \mu|_{B^c}(A) &\geq -\frac{\mu(B)}{1 - \mu(B)} \mu(A \cap B^c) \\ &\geq -\frac{\mu(B)}{1 - \mu(B)} \mu(B^c) \\ &= -\mu(B)\,,\end{aligned}$$

thus establishing the assertion. □

For $B \in \mathcal{B}(\mathcal{X})$ and $\mu \in \mathcal{M}(\mathcal{X})$ such that $\mu(B) < 1$, the conditional measure $\mu|_{B^c} \in \mathcal{M}(\mathcal{X})$ is defined by

$$\mu|_{B^c}(A) := \frac{\mu(A \cap B^c)}{\mu(B^c)}, \quad A \in \mathcal{B}(\mathcal{X}).$$

Consider the total variation metric d_{TV} on $\mathcal{M}(\mathcal{X})$ defined by

$$d_{\mathrm{TV}}(\mu_1, \mu_2) := \sup_{A \in \mathcal{B}(\mathcal{X})} |\mu_1(A) - \mu_2(A)|$$

It follows from Proposition 8.1 that $d_{\mathrm{TV}}(\mu, \mu|_{B^c}) \leq \mu(B)$ and since $d_{\mathcal{M}} \leq d_{\mathrm{TV}}$ (see e.g. [68, Eq. 2.24]), we conclude that

$$d_{\mathcal{M}}(\mu, \mu|_{B^c}) \leq \mu(B). \tag{8.11}$$

Let $B_\delta := B_\delta(x_1)$ denote the ball about the first sample of $x^n = (x_1, \ldots, x_n)$. Then, for $\mu_0 \in \mathcal{A}_0$, it follows from the assumptions that

$$\begin{aligned}d_{\mathcal{M}}(\mu_0, \mu_0|_{B_\delta^c}) &\leq \mu_0(B_\delta) \\ &\leq \mathcal{P}^\infty(\delta) \\ &< \alpha\end{aligned}$$

and therefore

$$\left(\mu_0, \mu_0|_{B_\delta^c}\right) \in \Psi^{-1}\mu_0.$$

Moreover, since

$$\begin{aligned}\mathbb{D}^n_{(\mu_0,\mu_0|_{B_\delta^c})}[B_\delta^n] &= \big(\mu_0|_{B_\delta^c}\big)^n[B_\delta^n] \\ &\le \mu_0|_{B_\delta^c}[B_\delta] \\ &= 0,\end{aligned}$$

we conclude that the condition (6.1)

$$\inf_{(\mu_0,\mu_0')\in\Psi^{-1}\mu_0} \mathbb{D}^n(\mu_0,\mu_0')[B_\delta^n] = 0$$

of Theorem 6.1 is satisfied for all $\mu_0 \in \mathcal{A}_0$.

Now consider the diagonal map $\Delta\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ defined by

$$\Delta(\mu) := (\mu,\mu), \quad \mu \in \mathcal{M}(\mathcal{X}).$$

Since

$$\Psi \circ \Delta(\mu) = P_0 \circ \Delta(\mu) = \mu, \quad \text{for all } \mu \in \mathcal{M}(\mathcal{X}),$$

it follows, if we define Δ on the first component of the product $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ and then restrict to $\mathcal{A}_0$, that Δ is a section of $\Psi = P_0$. It is clearly measurable, but also satisfies

$$P_\alpha \circ \Delta(\mu) = \mu, \quad \text{for all } \mu \in \mathcal{M}(\mathcal{X}),$$

that is, $P_\alpha \circ \Delta$ is the identity map from the first component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$ to the second. Then, for $\mu_0 \in \mathcal{A}_0$, the positivity of the model $\mathcal{P}$ implies that

$$\begin{aligned}\mathbb{D}^n\big(\Delta(\mu_0)\big)[B_\delta^n] &= \mathbb{D}_0^n \circ P_\alpha\big(\Delta(\mu_0)\big)[B_\delta^n] \\ &= \mathbb{D}_0^n(\mu_0)[B_\delta^n] \\ &= (\mu_0)^n[B_\delta^n] \\ &= \prod_{i=1}^n \mu_0[B_\delta(x_i)] \\ &> 0\end{aligned}$$

so that the second condition (6.2) of Theorem 6.1 is satisfied for all $\mu_0 \in \mathcal{A}_0$. Theorem 6.1 then asserts that

$$\mathcal{U}(\Psi^{-1}\Pi_0 \odot_{B_\delta^n} \mathbb{D}^n) \ge \Pi_0^\infty(\Phi \circ \Delta).$$

Moreover, since

$$\Phi \circ \Delta = \Phi_0 \circ P_\alpha \circ \Delta = \Phi_0,$$

now as a function on the first component of $\mathcal{M}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})$, and

$$\Psi^{-1}\Pi_0 = P_0^{-1}\Pi_0 = \Pi,$$

we conclude that

$$\mathcal{U}(\Pi \odot_{B_\delta^n} \mathbb{D}^n) \ge \Pi_0^\infty(\Phi_0).$$

The identity $\mathcal{U}(\Pi \odot_{B_\delta^n} \mathbb{D}^n) = \mathcal{U}(\Pi_\alpha \odot_{B_\delta^n} \mathbb{D}_0^n)$ of (8.10) then implies the assertion.

8.12 Proof of Theorem 7.1

By the discussion at the beginning of Section 7.2, $C(\mathcal{X})$ is a Polish evaluation measurable function space. By Theorem 7.10, such RKHSs or RKBSs are Polish evaluation measurable function spaces. Also, Theorem 7.12 proves that $\mathrm{UC}(\mathcal{X})$ is a Polish evaluation measurable function space. The assertion then follows from Corollary 7.7.

8.13 Proof of Lemma 7.2

By [71, Thm. 13.11, pg. 84] (see also [4, Thm. 4.59]), there exists a Polish topology τ^* on $\mathcal{X}$ such that

$$\mathcal{B}(\tau^*) = \mathcal{B}(\tau_{\mathcal{X}})$$

and

$$f\colon (X, \tau^*) \to (\mathcal{Y}, \tau_{\mathcal{Y}})$$

is continuous. Then, by [4, Thm. 15.14],

$$f_*\colon \mathcal{M}(\mathcal{B}(\tau^*)) \to \mathcal{M}(\mathcal{B}(\tau_{\mathcal{Y}}))$$

is continuous. Since $\mathcal{B}(\tau^*) = \mathcal{B}(\tau_{\mathcal{X}})$, it follows that $\mathcal{M}(\mathcal{B}(\tau^*)) = \mathcal{M}(\mathcal{B}(\tau_{\mathcal{X}}))$, and the result follows.

8.14 Proof of Proposition 7.3

The first assertion follows by observing that the product metric makes the product of complete spaces complete. The assertion that J is measurable follows from Theorem 7.5, the fact that point evaluation on $B(X)$ is continuous and therefore measurable, and that $\mathcal{B}(\mathcal{F}(\mathcal{X})) \times \mathcal{B}(\mathcal{M}(\mathcal{X})) \subseteq \mathcal{B}(\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X}))$.

8.15 Proof of Theorem 7.5

Let us begin by proving the "if" part. To prove measurability in the product structure, we would like to show that J is a Carathéodory function (see e.g. [4, Def. 4.50]), meaning that fixing f it is continuous and fixing μ it is measurable. In that case, since $\mathcal{X}$ is Polish, it follows that $\mathcal{M}(\mathcal{X})$ is Polish, and since $(\mathcal{F}, \sigma(\mathcal{F}))$ is measurable and $\mathcal{M}(\mathbb{R})$ is metrizable, it follows from [4, Lem. 4.51] that J is measurable.

To that end, observe that Lemma 7.2 implies that the map $f_*\colon \mathcal{M}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ is continuous, so it remains to show that, for $\mu \in \mathcal{M}(\mathcal{X})$, the map $\mu^t\colon \mathcal{F} \to \mathcal{M}(\mathbb{R})$ defined by

$$\mu^t(f) := f_*\mu, \quad f \in \mathcal{F},$$

is measurable. Consider a Dirac mass $\mu = \delta_x$ for $x \in \mathcal{X}$. Then,

$$\delta_x^t f = \delta_{f(x)}.$$

In order to prove that δ_x^t is measurable, we metrize the weak-$*$ topology on $\mathcal{M}(\mathbb{R})$, which by [48, Prop. 11.3.3] can be accomplished with the *bounded Lipschitz metric* b defined by

$$b(\nu_1, \nu_2) := \sup_{\|h\|_{\mathrm{BL}} \leq 1} \int_{\mathbb{R}} h \, \mathrm{d}(\nu_1 - \nu_2), \quad \nu_1, \nu_2 \in \mathcal{M}(\mathbb{R}), \tag{8.12}$$

defined in terms of any metric d on $\mathbb{R}$ that generates the same topology as the standard metric. By [48, Thm. 2.8.2], there is a totally bounded metrization e of $\mathbb{R}$ such that the completion with respect to that metric is compact. Choose such a metric e for the definition of the metric b in (8.12). Now the space $\mathrm{BL}(\mathbb{R}, e)$ of bounded Lipschitz functions on $(\mathbb{R}, e)$ is isomorphic to the bounded Lipschitz functions on the completion with respect to e. However, this completion space is compact, and the Arzelà–Ascoli Theorem [7, Thm. A8.5] implies that the unit ball in the space of bounded Lipschitz functions on the completion is compact. Consequently, it follows that the unit ball in $\mathrm{BL}(\mathbb{R}, e)$ is also compact. With this metrization, [4, Lem. 4.30] implies that to prove that δ_x^t is measurable it is sufficient to prove that

$$f \mapsto b(\delta_x^t f, \mu)$$

is measurable for all $\mu \in \mathcal{M}(\mathbb{R})$.

Consequently, the compactness of the unit ball of $\mathrm{BL}(\mathbb{R}, e)$ and the continuity of $\nu \colon B(\mathbb{R}) \to \mathbb{R}$ defined by $\nu(h) := \int_{\mathbb{R}} h \, \mathrm{d}\nu$ for all $\nu \in \mathcal{M}(\mathbb{R})$ together imply that there exists a countable subset $\{h_i \mid i \in \mathbb{N}\}$ of the unit ball of $\mathrm{BL}(\mathbb{R}, e)$ such that

$$\begin{aligned} b(\delta_x^t f, \mu) &:= \sup_{i \in \mathbb{N}} \int_{\mathbb{R}} h_i \mathrm{d}(\delta_x^t f - \mu_2) \\ &= \sup_{i \in \mathbb{N}} \int_{\mathbb{R}} h_i \, \mathrm{d}(\delta_{f(x)} - \mu_2) \\ &= \sup_{i \in \mathbb{N}} \left(h_i\big(f(x)\big) - \int_{\mathbb{R}} h_i \, \mathrm{d}\mu_2 \right). \end{aligned}$$

Consequently, if each function

$$f \mapsto h_i\big(f(x)\big) \tag{8.13}$$

is measurable then it follows from [48, Thm. 4.2.2] — which states that the pointwise limit of a convergent sequence of measurable functions is measurable — that δ_x^t is measurable. However, the assumption that $(\mathcal{F}, \sigma(\mathcal{F}))$ is evaluation measurable, the measurability of h, and the fact that the composition of measurable functions is measurable implies that the function (8.13) is measurable. Therefore, δ_x^t is measurable for each $x \in \mathcal{X}$. Hence, μ^t is measurable for all measures μ with finite support. Since, by [4, Thm. 15.10], the measures with finite support are dense in $\mathcal{M}(\mathcal{X})$, the continuity as a function of μ for fixed f (assured by Lemma 7.2) implies that μ^t, for arbitrary $\mu \in \mathcal{M}(\mathcal{X})$, is a pointwise convergent limit of a sequence of measurable functions into a metric space, and therefore by [48, Thm. 4.2.2] is measurable. Therefore, J is a Carathéodory function and the "if" part of the assertion is established.

For the "only if" part, observe that, if J is measurable in the product structure, then, for $x \in \mathcal{X}$ and the corresponding Dirac mass $\delta_x \in \mathcal{M}(\mathcal{X})$, it follows that $J(\,\cdot\,, \delta_x)\colon \mathcal{F}(\mathcal{X}) \to \mathcal{M}(\mathbb{R})$ is measurable. By [4, Thm. 15.8], the map $r \mapsto \delta_r$ is an embedding of $\mathbb{R}$ into $\mathcal{M}(\mathbb{R})$. Then, since $J(f, \delta_x) = \delta_{f(x)}$, composition with the inverse of this embedding yields that $f \mapsto f(x)$ is measurable for all $x \in \mathcal{X}$. That is, $(\mathcal{F}(\mathcal{X}), \sigma(\mathcal{F}))$ is an evaluation measurable function space.

8.16 Proof of Corollary 7.6

By Theorem 7.5, J is measurable, and, by Lemma 7.2, $h_*\colon \mathcal{M}(\mathbb{R}) \to \mathcal{M}(\mathbb{R})$ is continuous. Therefore, since

$$J_h(f, \mu) = (h \circ f)_* \mu = h_* f_* \mu = h_* J(f, \mu),$$

it follows that $J_h = h_* \circ J$ and the assertion follows.

8.17 Proof of Corollary 7.7

Since $\mathcal{X}$ is Polish, it follows that $\mathcal{M}(\mathcal{X})$ is Polish, and, since $(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F}))$ is Polish, it follows from [48, Prop. 2.1.4] that both $(\mathcal{F}(\mathcal{X}), \tau(\mathcal{F}))$ and $\mathcal{M}(\mathcal{X})$ are second countable. Therefore, by [48, Prop. 4.1.7], the product of the Borel structures is the Borel structure of the product. That is,

$$\mathcal{B}\big(\tau(\mathcal{F})\big) \times \mathcal{B}\big(\mathcal{M}(\mathcal{X})\big) = \mathcal{B}\big(\mathcal{F}(\mathcal{X}) \times \mathcal{M}(\mathcal{X})\big).$$

The assertion then follows from Theorem 7.5.

8.18 Proof of Lemma 7.8

Let us begin with the "if" part. To that end, let us first show that condition (7.4) implies that $\Phi(\mathcal{X})$ is separable. Indeed, fix $\varepsilon > 0$ and for each $\frac{\varepsilon}{2^k}$, $k \in \mathbb{N}$, let B_j^k, $j \in \mathbb{N}$, denote the corresponding partition and let $x_j^k \in B_j^k$ denote a selection. The set

$$\{\Phi(x_j^k) \mid k \in \mathbb{N}, j \in \mathbb{N}\}$$

is countable, and it is easy to show it is dense in $\Phi(\mathcal{X})$. Hence, $\Phi(\mathcal{X})$ is separable. Consequently, $\operatorname{span}\{\Phi(\mathcal{X})\}$ is separable, and, therefore, $\mathcal{W} := \overline{\operatorname{span}}\{\Phi(\mathcal{X})\}$ is separable. It then follows that $B = \{\langle u, \Phi^*(\,\cdot\,)\rangle\}$ with norm $\|\langle u, \Phi^*(\cdot)\rangle\| = \|u\|_{\mathcal{W}}$ is separable.

For the "only if" part, suppose that B is separable. Then the feature space $\mathcal{W} := B$ is separable and, since B is metric, it follows for the corresponding feature map $\Phi\colon \mathcal{X} \to B$ that $\Phi(\mathcal{X}) \subseteq B$ is separable. Therefore, there exists a countable dense set $\{\Phi(x_j) \mid j \in \mathbb{N}\}$ of $\Phi(\mathcal{X})$. Therefore, if for each $\varepsilon > 0$ and for each $j \in \mathbb{N}$ we define

$$B_j := \left\{ x \in \mathcal{X} \,\middle|\, \|\Phi(x_j) - \Phi(x)\|_B < \frac{\varepsilon}{2} \right\},$$

it follows that $\bigcup_{j \in \mathbb{N}} B_j = \mathcal{X}$ and $\|\Phi(x_1) - \Phi(x_2)\|_B < \varepsilon$ for all $x_1, x_2 \in B_j$.

8.19 Proof of Lemma 7.9

As in the proof of Lemma 7.4, it is sufficient to prove that the the range of the feature map is separable. First recall that cite[Thm. 2]Frolik:1963 has shown that bianalyticity is equivalent to being separable and absolutely Borel. Then, observe that Stone's Theorem [106, Thm. 16, pg. 32] states that when $\mathcal{X}$ is a separable absolutely Borel space and $\Phi\colon \mathcal{X} \to \mathcal{Y}$ is a measurable bijection between $\mathcal{X}$ and a metric space $\mathcal{Y}$, then the image $\Phi(\mathcal{X}) = \mathcal{Y}$ is separable. However, bijectivity of Φ is in fact not necessary. To see this, select a metric on $\mathcal{X}$ that generates its topology and extend to the injective map $\hat{\Phi}\colon \mathcal{X} \to \mathcal{X} \times \mathcal{Y}$ defined by $\hat{\Phi}(x) := (x, \Phi(x))$, where, in the product, $\hat{\Phi}$ is measurable. Then, by restricting to the range, with its inherited metric, we obtain that the resulting $\hat{\Phi}$ is bijective, and so the range of $\hat{\Phi}$ is separable by Stone's Theorem. Since the range of Φ is the continuous image of the range of $\hat{\Phi}$ and separability is preserved by continuous maps, the range of Φ is separable.

8.20 Proof of Theorem 7.10

Theorem 12 of [57] shows that $\mathcal{X}$ is absolutel Borel when it is Polish, and therefore it is also bianalytic. Consequently, Lemma 7.9 implies the separability of $\mathcal{K}$. Since Banach and Hilbert spaces are complete metric spaces, it follows that $\mathcal{K}$ is Polish. Moreover, the measurability of a feature map for a RKHS or the measurability of a dual feature map for the RKBS guarantees that the space consists of measurable functions. Since point evaluation is continuous it is measurable, which completes the proof.

8.21 Proof of Theorem 7.12

Let us first show that $\big(\mathrm{UC}(\mathcal{X}), \mathcal{B}\big(\tau_W(\mathrm{UC})\big)\big)$ is Polish. To that end, recall that Beer [15] has shown that $\big(\mathrm{CL}(\mathcal{X} \times \mathbb{R}), \tau_W\big)$ is Polish. Therefore, if $\mathrm{hypo}\big(UC(\mathcal{X})\big) \subset CL(\mathcal{X} \times \mathbb{R})$ is closed, it follows that $\mathrm{hypo}\big(\mathrm{UC}(\mathcal{X})\big)$ is Polish. Since hypo is one-to-one, it follows from [32] that with the pullback topology $\tau_W(\mathrm{UC})$ hypo is an embedding, and therefore $\big(\mathrm{UC}(\mathcal{X}), \mathcal{B}\big(\tau_W(\mathrm{UC})\big)\big)$ is Polish. So let us show that $\mathrm{hypo}\big(\mathrm{UC}(\mathcal{X})\big)$ is closed. Let $\mathrm{hypo}(f_i) \to A$ and suppose that A is not a hypograph. Then there exist $x \in \mathcal{X}$ and $r_1 < r_2 < r_3$ such that $(x, r_1) \in A$, $(x, r_2) \notin A$ and $(x, r_k) \in A$. Consequently, by [16, Proposition, pg. 80] (which Beer credits to Del Prete and Lignola [40]), it follows that if we select small enough non overlapping open cylinders B_1, B_2, B_3 with the same base about $(x, r_1), (x, r_2), (x, r_{i3})$ respectively, then there exists $N \in \mathbb{N}$ such that that, for all $n \geq N$, $\mathrm{hypo}(f_n) \cap B_1 \neq \varnothing$, $\mathrm{hypo}(f_n) \cap B_2 = \varnothing$, and $\mathrm{hypo}(f_n) \cap B_3 \neq \varnothing$. Since the bases of the cylinders are the same, this contradicts the fact that $\mathrm{hypo}(f_n)$ is a hypograph. Therefore, A is a hypograph and so $\mathrm{hypo}\big(\mathrm{UC}(\mathcal{X})\big) \subset CL(\mathcal{X} \times \mathbb{R})$ is closed and we have proved that $hypo\big(UC(\mathcal{X})\big)$ is Polish.

Now let us show it is an evaluation measurable function space. It follows from the

proof of [120, Thm. 6.1] that if $hypo(f_n) \to hypo(f)$ then we have

$$\begin{aligned} f &= \lim_{\rho} \limsup_{n \mapsto \infty} {}_{\rho} f_n \\ &\geq \lim_{\rho} \limsup_{n \mapsto \infty} f_n \\ &= \limsup_{n \mapsto \infty} f_n \end{aligned}$$

and, therefore, for all $x \in \mathcal{X}$,

$$f(x) \geq \limsup_{n \mapsto \infty} f_n(x)$$

or written in another way

$$i_x(f) \geq \limsup_{n \mapsto \infty} i_x(f_n)$$

where i_x is point evaluation. Therefore it follows from the alternative characterization of upper semicontinuity [4, Lem. 2.42] that point evaluation is upper semicontinuous in the pull-back topology $\tau_W(\mathrm{UC})$ and, therefore, measurable with respect to the corresponding Borel σ-algebra.

9 Appendix

The following lemma is Lemma III.39 p. 86 of [34]. We also refer to p. 87 of [34] for the existence of the measurable selection η (which is also derived from Theorem III.38 p.85 of [34]). These results are related to Aumann's measurable section principle [8] (the extension to Suslin space is due to Sainte-Beuve [94]).

Lemma 9.1. *Let $(T, \mathcal{T})$ be a measurable space, S a Suslin space. $\varphi\colon T \times S \to \bar{R}$ a $\mathcal{T} \otimes \mathcal{B}(S)$ measurable function and Γ a multifunction (i.e. a set-valued map) from T to non-empty subsets of S whose graph G belongs to $\mathcal{T} \times \mathcal{B}(S)$. Then*

1. *the function*
$$m(t) := \sup\{\phi(t, x) \mid x \in \Gamma(t)\}$$
is a $\hat{\mathcal{T}}$ measurable function of t.

2. *for $\delta > 0$, there exists η, a $\hat{\mathcal{T}}$ measurable function of t, such that $\eta(t) \in \Gamma(t)$ and $\varphi(t, \eta(t)) > m(t) - \delta$.*

The following definition is Definition 4.50 in [4]:

Definition 9.2. Let (S, Σ) be a measurable space, and let X and Y be topological spaces. A function $h\colon S \times X \to Y$ is a *Carathéodory function* if:

1. for each $x \in X$, the function $h^x = h(., x) : S \to Y$ is $\big(\Sigma, \mathcal{B}(Y)\big)$-measurable; and

2. for each $s \in S$, the function $h_s = h(s, .) : X \to Y$ is continuous.

The following lemma is Lemma 4.51 in [4] (see also [34, p. 70]):

Lemma 9.3. *Let (S, Σ) be a measurable space, X a separable metrizable space, and Y a metrizable space. Then every Carathéodory function $h\colon S \times X \to Y$ is jointly measurable.*

9.1 Universally measurable functions

For a topological space T let $\hat{\mathcal{B}}(T)$ denote the σ-algebra of universally measurable sets. For a measure μ, let $\hat{\mu}$ denote its completion. Here we state the following proposition that allows us to define the expected value of $\hat{\mathcal{B}}(T)$ measurable functions with respect to Borel measures. In all statements in the following proposition, the assertions follow when the integrals involved exist, in particular for semibounded functions. The proof is straightforward but tedious and follows from e.g. [45, Thm. pg. 37], the English version of [41, Ch. 2, pg. 49], and [34].

Proposition 9.4. *Let T be a topological space. Then we have*

- *For a measurable function f we have $\mathbb{E}_{\hat{\mu}} f = \mathbb{E}_{\mu} f$*
- *Let f be $\hat{\mathcal{B}}(T)$-measurable. Then there exist two measurable functions $\underline{f}$ and $\overline{f}$ such that*
$$\underline{f} \leq f \leq \overline{f}, \quad \mu(\underline{f} \neq \overline{f}) = 0$$
and, for any such functions, we have
$$\mathbb{E}_{\mu}[\underline{f}] = \mathbb{E}_{\hat{\mu}}[f] = \mathbb{E}_{\mu}[\overline{f}]$$
- *For a fixed μ, $f \mapsto \mathbb{E}_{\hat{\mu}}[f]$ defines an affine function on the cone of non-negative $\hat{\mathcal{B}}(T)$-measurable functions*
- *For a fixed $\hat{\mathcal{B}}(T)$-measurable function f, the function $\mathcal{M}(T) \ni \mu \mapsto \mathbb{E}_{\hat{\mu}}[f]$ is affine.*
- *Suppose that f_1, f_2 are $\hat{\mathcal{B}}(T)$-measurable non-negative functions such that $f_1 \leq f_2$. Then $\mathbb{E}_{\hat{\mu}}[f_1] \leq E_{\hat{\mu}}[f_2]$ for all $\mu \in \mathcal{M}(T)$.*

Proposition 9.4 leads to the following definition for the expectation of $\hat{\mathcal{B}}(T)$-measurable functions with respect to Borel probability measures on T:

Definition 9.5. For a Borel probability measure $\mu \in \mathcal{M}(T)$, we define the integral of a $\hat{\mathcal{B}}(T)$-measurable function f by
$$\mathbb{E}_{\mu}[f] := \mathbb{E}_{\hat{\mu}}[f]$$
when the latter exists, where $\hat{\mu}$ is the completion of the measure μ as described in [45, p. 37].

Recall that a *carrier* T for a probability measure $\mathbb{Q} \in \mathcal{M}(\mathcal{Q})$ is a set $T \in \mathcal{B}(\mathcal{Q})$ such that $\mathbb{Q}(T) = 1$. For a carrier T, since $T \in \mathcal{B}(\mathcal{Q})$, it follows that $\mathcal{B}(T) = \mathcal{B}(\mathcal{Q}) \cap T$ and we can define the trace measure $\mathbb{Q}_T \in \mathcal{M}(T)$ by $\mathbb{Q}_T(A) := \mathbb{Q}(A), A \in \mathcal{B}(\mathcal{Q}) \cap T$. The following proposition shows that the expectation of a function can be defined with respect to measures which possess carriers upon which the function is universally measurable:

Proposition 9.6. *Let S be a topological space. Suppose that f is $\hat{\mathcal{B}}(T)$-measurable for all measurable $T \subseteq S$. Suppose also that $\mathbb{Q} \in \mathcal{M}(S)$ has a carrier $T \subseteq S$. Then, using Definition 9.5, any such carrier T defines an expectation*

$$\mathbb{E}_{\mathbb{Q}_T}[f] := \mathbb{E}_{\hat{\mathbb{Q}}_T}[f],$$

and this definition is independent of the carrier; that is, if $T' \subset S$ is another carrier, then

$$\mathbb{E}_{\hat{\mathbb{Q}}_{T'}}[f] = \mathbb{E}_{\hat{\mathbb{Q}}_T}[f].$$

Moreover, this expectation satisfies the assertions of affinity and monotonicity of Proposition 9.4.

We also need a change of variables formula for expectations of universally measurable functions.

Proposition 9.7. *Let X and Y be topological spaces, $\Psi\colon X \to Y$ a measurable map and suppose that $f\colon Y \to \mathbb{R}$ is $\hat{\mathcal{B}}(Y)$ measurable. Then $f \circ \Psi\colon X \to \mathbb{R}$ is $\hat{\mathcal{B}}(X)$-measurable and, for $\pi \in \mathcal{M}(X)$,*

$$\mathbb{E}_{\Psi\pi}[f] = \mathbb{E}_{\pi}[f \circ \Psi].$$

For Suslin space $\mathcal{X}$ and a subset $M \subset \mathcal{M}(\mathcal{X})$ let $\Sigma(M)$ denote the smallest σ-subalgebra of subsets of M for which the the evaluation map $\nu \mapsto \nu(B)$ is measurable for all $B \in \mathcal{B}(\mathcal{X})$. The following version of a result of Weizsacker and Winkler [111] as stated in [121, Thm. 3.1] will be useful to us:

Theorem 9.8. *Consider a Suslin space $\mathcal{X}$, n real valued measurable functions $f_i\colon \mathcal{X} \to R$, n constants $c_1, \ldots, c_n \in \mathbb{R}$, and define*

$$H := \left\{\nu \in \mathcal{M}(\mathcal{X}) \;\middle|\; f_i \text{ is } \nu\text{-integrable and } \mathbb{E}_\nu[f_i] \leq c_i, \text{ for } i = 1, \ldots, n\right\}$$

Then, for each $\nu \in H$, there is a probability measure p on $\Sigma(\mathrm{ext}(H))$ such that

$$\nu(B) = \int_{\mathrm{ext}(H)} \nu'(B)\, \mathrm{d}p(\nu'), \quad \text{for all } B \in \mathcal{B}(\mathcal{X}). \tag{9.1}$$

[121, Prop. 3.1] shows that if a measurable function $f\colon \mathcal{X} \to \mathbb{R}$ is integrable with respect to all measures in H (allowing the values ∞ and $-\infty$), then integration

$$F(\nu) := \int_{\mathcal{X}} f \,\mathrm{d}\nu$$

is measure affine per Definition 4.3. We need a slightly more general result:

Lemma 9.9. *Consider the situation of Theorem 9.8, let $f\colon \mathcal{X} \to \mathbb{R}$ be a semibounded universally measurable function. Then*

$$F(\nu) := \mathbb{E}_{\hat{\nu}}[f], \quad \text{for } \nu \in H,$$

is measure affine per Definition 4.3.

The next lemma extends [121, Thm. 2.1] to the case where the constraint functions f_i, for $i = 1, \ldots, n$, are universally measurable:

Lemma 9.10. *Let $\mathcal{X}$ be Suslin, and fix universally measurable real-valued functions $f_1, \ldots, f_n$ and constants $c_1, \ldots, c_n$. Then*

$$H := \left\{ \nu \in \mathcal{M}(\mathcal{X}) \,\middle|\, f_i \text{ is } \hat{\nu}\text{-integrable and } \mathbb{E}_{\hat{\nu}}[f_i] \leq c_i \text{ for } i = 1, \ldots, n \right\} \tag{9.2}$$

is convex and

$$\operatorname{ext}(H) = \left\{ \nu \in H \,\middle|\, \nu = \sum_{i=1}^{m} \alpha_i \delta_{x_i}, \alpha_i \geq 0, x_i \in \mathcal{X}, i = 1, \ldots, m, \sum_{i=1}^{m} \alpha_i = 1, 1 \leq m \leq n+1, \right.$$
$$\left. \text{the vectors} \big(f_1(x_i), f_2(x_i), \ldots, f_n(x_i), 1\big), 1 \leq i \leq m \text{ are linearly independent} \right\}.$$

9.2 Proofs

9.2.1 Proof of Proposition 9.6

Let T and T' be two carriers for $\mathbb{Q} \in \mathcal{M}(S)$ and f a function such that f_T and $f_{T'}$ are $\hat{\mathcal{B}}(t)$ and $\hat{\mathcal{B}}(T')$ measurable respectively. Then Proposition 9.4 implies that there are functions f_1, f_2 measurable on T and f_1', f_2' measurable on T' such that

$$\begin{aligned} f_1 \leq f_T \leq f_2 \qquad & \mathbb{Q}_T(f_1 \neq f_2) = 0 \\ f_1' \leq f_{T'} \leq f_2' \qquad & \mathbb{Q}_{T'}(f_1' \neq f_2') = 0 \end{aligned}$$

so that

$$\begin{aligned} \mathbb{E}_{\hat{\mathbb{Q}}_T}[f_T] &= \mathbb{E}_{\mathbb{Q}_T}[f_1] \\ \mathbb{E}_{\hat{\mathbb{Q}}_{T'}}[f_{T'}] &= \mathbb{E}_{\mathbb{Q}_{T'}}[f_1'] \end{aligned}$$

Now, it is easy to see that $T \cap T'$ is also a carrier and that we have

$$f_1(x) \leq f(x) \leq f_2(x), \quad x \in T \cap T'$$

and

$$\mathbb{Q}_{T \cap T'}(f_1 \neq f_2) \leq \mathbb{Q}_T(f_1 \neq f_2) = 0$$

so that we conclude from Proposition 9.4 that

$$\begin{aligned} \mathbb{E}_{\hat{\mathbb{Q}}_{T \cap T'}}[f] &= \mathbb{E}_{\mathbb{Q}_{T \cap T'}}[f_1] \\ &= \mathbb{E}_{\mathbb{Q}_T}[f_1] - \mathbb{E}_{\mathbb{Q}_{T \setminus T'}}[f_1] \\ &= \mathbb{E}_{\mathbb{Q}_T}[f_1] \\ &= \mathbb{E}_{\hat{\mathbb{Q}}_T}[f] \end{aligned}$$

and so conclude that

$$\mathbb{E}_{\hat{\mathbb{Q}}_{T \cap T'}}[f] = \mathbb{E}_{\hat{\mathbb{Q}}_T}[f].$$

By the same argument on T' we conclude that $\mathbb{E}_{\hat{\mathbb{Q}}_{T \cap T'}}[f] = \mathbb{E}_{\hat{\mathbb{Q}}_{T'}}[f]$ and therefore the first assertion is proved. The assertions of affinity and monotonicity are similarly straightforward.

9.2.2 Proof of Proposition 9.7

Consider $\pi \in \mathcal{M}(X)$ and its pushforward $\nu := \Psi\pi$. By Proposition 9.4 and the assumptions, there exists two measurable functions $\underline{f}$ and $\overline{f}$ such that

$$\underline{f} \leq f \leq \overline{f}, \quad \nu(\underline{f} \neq \overline{f}) = 0$$

from which we conclude that

$$\underline{f} \circ \Psi \leq f \circ \Psi \leq \overline{f} \circ \Psi$$

and

$$\begin{aligned} 0 &= \nu[\underline{f} \neq \overline{f}] \\ &= \Psi\pi[\underline{f} \neq \overline{f}] \\ &= \pi[\Psi^{-1}\{\underline{f} \neq \overline{f}\}] \\ &= \pi[\underline{f} \circ \Psi \neq \overline{f} \circ \Psi] \end{aligned}$$

so that we obtain

$$\pi[\underline{f} \circ \Psi \neq \overline{f} \circ \Psi] = 0\,.$$

Since π was arbitrary, it follows that $f \circ \Psi$ is $\hat{\mathcal{B}}(X)$-measurable. To obtain the change of variables formula, compute

$$\begin{aligned} \mathbb{E}_\pi[f \circ \Psi] &:= \mathbb{E}_{\hat{\pi}}[f \circ \Psi] \\ &= \mathbb{E}_\pi[\overline{f} \circ \Psi] \\ &= \mathbb{E}_{\Psi\pi}[\overline{f}] \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}_{\Psi\pi}[f] &:= \mathbb{E}_{\widehat{\Psi\pi}}[f] \\ &= \mathbb{E}_{\Psi\pi}[\overline{f}] \end{aligned}$$

from which we conclude the change of variables formula

$$\mathbb{E}_\pi[f \circ \Psi] = \mathbb{E}_{\Psi\pi}[f].$$

9.2.3 Proof of Lemma 9.9

Fix $\nu \in H$ and a probability measure p such that the barycentric formula (9.1) holds. Proposition 9.4 asserts that there are measurable functions $f_1 \leq f \leq f_2$ such that $\nu(f_1 \neq f_2) = 0$. Therefore, $f_2 - f \geq 0$, $\mathbb{E}_{\hat{\nu}}(f_2 - f) = 0$, $f - f_1 \geq 0$, and $\mathbb{E}_{\hat{\nu}}(f - f_1) = 0$. Moreover, it is easy to see then we can make both f_1 and f_2 semibounded. Therefore F is a well defined extended real valued function. Moreover, [121, Prop. 3.1] asserts that the function $\nu \mapsto \mathbb{E}_\nu[f_i]$ is measure affine for $i = 1, 2$, and so

$$\mathbb{E}_\nu[f_i] = \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_i]\, \mathrm{d}p(\nu'), \quad \text{for } i = 1, 2.$$

Consequently, since $\nu[f_1 \neq f_2] = 0$, it follows that $\mathbb{E}_\nu[f_2 - f_1] = 0$ so that

$$0 = \mathbb{E}_\nu[f_2 - f_1] = \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_2 - f_1] \, \mathrm{d}p(\nu'), \quad \text{for } i = 1, 2,$$

and since $f_2 - f_1 \geq 0$ it follows that

$$\nu'[f_2 \neq f_1] = 0, \quad p\text{-a.e.}$$

and therefore

$$\hat{\nu'}[f \neq f_1] = 0, \quad p\text{-a.e.}$$

Therefore we conclude that

$$\begin{aligned}
F(\nu) &:= \mathbb{E}_{\hat{\nu}}[f] \\
&= \mathbb{E}_{\hat{\nu}}[f_1] + \mathbb{E}_{\hat{\nu}}[f - f_1] \\
&= \mathbb{E}_{\hat{\nu}}[f_1] \\
&= \mathbb{E}_{\nu}[f_1] \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\nu'}[f_1] \, \mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\hat{\nu'}}[f_1] \, \mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\hat{\nu'}}[f_1] \, \mathrm{d}p(\nu') + \int_{\mathrm{ext}(H)} \mathbb{E}_{\hat{\nu'}}[f - f_1] \, \mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} \mathbb{E}_{\hat{\nu'}}[f] \, \mathrm{d}p(\nu') \\
&= \int_{\mathrm{ext}(H)} F(\nu') \, \mathrm{d}p(\nu'),
\end{aligned}$$

and the assertion is proved.

9.2.4 Proof of Lemma 9.10

Let us first establish that

$$\widehat{\nu_1 + \nu_2} = \hat{\nu}_1 + \hat{\nu}_2, \quad \text{for all } \nu_1, \nu_2 \in \mathcal{M}(\mathcal{X}), \tag{9.3}$$

$$\widehat{\alpha\nu} = \alpha\hat{\nu}, \quad \text{for all } \nu \in \mathcal{M}(\mathcal{X}). \tag{9.4}$$

This follows from the fact that $(\nu_1 + \nu_2)(N) = 0$ if and only if $\nu_j(N) = 0$ for $j = 1, 2$ and the characterization of the completion $\hat{\nu}$ by

$$\hat{\nu}(B \cup S) := \nu(B), \quad B \in \mathcal{B}(\mathcal{X}),\, S \subset N,\, \nu(N) = 0$$

as found, for example, in [7, p. 18]. For then, for such B and S,

$$\begin{aligned}\widehat{\nu_1+\nu_2}(B\cup S) &= (\nu_1+\nu_2)(B)\\ &= \nu_1(B)+\nu_2(B)\\ &= \widehat{\nu_1}(B\cup S)+\widehat{\nu_2}(B\cup S)\end{aligned}$$

Now for the proof of the main assertion. Following the proof of [121, Thm. 2.1], it is sufficient to show that for

$$K := \big\{\nu\in\mathcal{M}(\mathcal{X})\,\big|\, f_i \text{ is } \hat{\nu}\text{-integrable for } i=1,\dots,n\big\},$$

we have

$$\operatorname{ext}(K) := \{\delta_x, x\in\mathcal{X}\}, \tag{9.5}$$

and that $\mathbb{R}_+K\subset\mathbb{R}_+\mathcal{M}(\mathcal{X})$ is a lattice cone in its own ordering. For the first, observe that since $\operatorname{ext}\big(\mathcal{M}(\mathcal{X})\big)=\{\delta_x\mid x\in\mathcal{X}\}$ and that f_i are δ_x-integrable for all $i=1,\dots,n$, $x\in\mathcal{X}$, it follows that

$$\{\delta_x\mid x\in\mathcal{X}\}\subseteq\operatorname{ext}(K).$$

Now suppose that $\nu\in\operatorname{ext}(K)$ is not a Dirac mass. Then, as in the proof that the extreme points of $\mathcal{M}(\mathcal{X})$ are the Dirac masses, see e.g. [4, Thm. 15.9], and using the fact that the support of ν must contain 2 or more points, we can decompose $\nu=\alpha\nu_1+(1-\alpha)\nu_2$ where $\nu_1\neq\nu_2$ and $\alpha\in(0,1)$. Moreover, from

$$\hat{\nu}=\alpha\hat{\nu_1}+(1-\alpha)\hat{\nu_2}$$

we conclude that f_i being $\hat{\nu}$-integrable implies that f_i is $\hat{\nu_j}$-integrable for $j=1,2$ and $i=1,\dots,n$. Consequently, $\nu_j\in K$ for $j=1,2$. Since ν was an extreme point we conclude that $\nu_1=\nu_2$ which is a contradiction, and (9.5) follows.

Now let us demonstrate that $\mathbb{R}_+K$ is a lattice cone in its own ordering. To that end, note that by [89, Lem. 10.4], it suffices to show that $\mathbb{R}_+K\subset\mathbb{R}_+\mathcal{M}(\mathcal{X})$ is a hereditary subcone, in that $\nu_1\in\mathbb{R}_+K$, $\nu_2\in\mathbb{R}_+\mathcal{M}(\mathcal{X})$ and $\nu_1-\nu_2\in\mathbb{R}_+K$ together imply that $\nu_2\in\mathbb{R}_+K$. To that end, consider such ν_1 and ν_2. Then (9.3) implies that $\widehat{(\nu_1-\nu_2)}=\hat{\nu}_1-\hat{\nu}_2$ and so we conclude that

$$0\le\mathbb{E}_{\widehat{(\nu_1-\nu_2)}}[|f_i|]=\mathbb{E}_{\hat{\nu_1}}[|f_i|]-\mathbb{E}_{\hat{\nu_2}}[|f_i|]$$

and therefore

$$\mathbb{E}_{\hat{\nu_2}}[|f_i|]\le\mathbb{E}_{\hat{\nu_1}}[|f_i|]<\infty,$$

from which we conclude that $\nu_2\in\mathbb{R}_+K$. Hence, $\mathbb{R}_+K$ is a hereditary subcone, and the assertion then follows as in the proof of [121, Thm. 2.1].

Acknowledgements

The authors gratefully acknowledge this work supported by the Air Force Office of Scientific Research under Award Number FA9550-12-1-0389 (Scientific Computation of Optimal Statistical Estimators).

References

[1] C. Abraham and B. Cadre. Asymptotic properties of posterior distributions derived from misspecified models. *C. R. Math. Acad. Sci. Paris*, 335(5):495–498, 2002. http://dx.doi.org/10.1016/S1631-073X(02)02520-7.

[2] C. Abraham and B. Cadre. Concentration of posterior distributions with misspecified models. *Ann. I.S.U.P.*, 52(3):3–14, 2008.

[3] F. Agterberg. Georges Matheron-founder of spatial statistics. pages 1–6, 2005. http://www.geostatscam.com/Adobe/G_Matheron.pdf.

[4] C. D. Aliprantis and K. C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, Berlin, third edition, 2006.

[5] R. F. Arens. A topology for spaces of transformations. *Ann. of Math. (2)*, 47:480–495, 1946.

[6] W. Arveson. *An Invitation to C*-Algebras*. Springer-Verlag, New York, 1976.

[7] R. B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972. Probability and Mathematical Statistics, No. 11.

[8] R. J. Aumann. Measurable utility and measurable choice theorem. *La décision C.N.R.S.*, pages 15–26, 1967.

[9] V. Bargmann. On a Hilbert space of analytic functions and an associated integral transform. *Comm. Pure Appl. Math.*, 14:187–214, 1961.

[10] A. Barron, M. J. Schervish, and L. Wasserman. The consistency of posterior distributions in nonparametric problems. *Ann. Statist.*, 27(2):536–561, 1999. http://dx.doi.org/10.1214/aos/1018031206.

[11] H. Bauer. *Measure and Integration Theory*, volume 26 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 2001. Translated from the German by Robert B. Burckel.

[12] G. Beer. A natural topology for upper semicontinuous functions and a Baire category dual for convergence in measure. *Pacific J. Math.*, 96(2):251–263, 1981.

[13] G. Beer. Metric spaces with nice closed balls and distance functions for closed sets. *Bull. Austral. Math. Soc.*, 35(1):81–96, 1987. http://dx.doi.org/10.1017/S000497270001306X.

[14] G. Beer. An embedding theorem for the Fell topology. *Michigan Math. J.*, 35(1):3–9, 1988. http://dx.doi.org/10.1307/mmj/1029003677.

[15] G. Beer. A Polish topology for the closed subsets of a Polish space. *Proc. Amer. Math. Soc.*, 113(4):1123–1133, 1991. http://dx.doi.org/10.2307/2048792.

[16] G. Beer. Wijsman convergence: a survey. *Set-Valued Anal.*, 2(1-2):77–94, 1994. Set convergence in nonlinear analysis and optimization.

[17] G. Beer, A. Lechicki, S. Levi, and S. Naimpally. Distance functionals and suprema of hyperspace topologies. *Ann. Mat. Pura Appl. (4)*, 162:367–381, 1992. http://dx.doi.org/10.1007/BF01760016.

[18] J. Berger. The case for objective Bayesian analysis. *Bayesian Anal.*, 1(3):385–402, 2006.

[19] J. O. Berger. The robust Bayesian viewpoint. In *Robustness of Bayesian Analyses*, volume 4 of *Stud. Bayesian Econometrics*, pages 63–144. North-Holland, Amsterdam, 1984. With comments and with a reply by the author.

[20] J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994. With comments and a rejoinder by the author.

[21] R. H. Berk. Limiting behavior of posterior distributions when the model is incorrect. *Ann. Math. Statist. 37 (1966), 51–58; correction, ibid*, 37:745–746, 1966.

[22] R. H. Berk. Consistency a posteriori. *Ann. Math. Statist.*, 41:894–906, 1970.

[23] A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Boston, MA, 2004. http://dx.doi.org/10.1007/978-1-4419-9096-9.

[24] S. N. Bernšteĭn. *Sobranie sochinenii. Tom IV: Teoriya veroyatnostei. Matematicheskaya statistika. 1911–1946*. Izdat. "Nauka", Moscow, 1964.

[25] D. Bertsimas and I. Popescu. Optimal inequalities in probability theory: a convex optimization approach. *SIAM J. Optim.*, 15(3):780–804 (electronic), 2005. http://dx.doi.org/10.1137/S1052623401399903.

[26] D. M. Blei, M. I. Jordan, and A. Y. Ng. Hierarchical Bayesian models for applications in information retrieval. In *Bayesian statistics, 7 (Tenerife, 2002)*, pages 25–43. Oxford Univ. Press, New York, 2003.

[27] V. I. Bogachev. *Measure Theory. Vol. II*. Springer-Verlag, Berlin, 2007. http://dx.doi.org/10.1007/978-3-540-34514-5.

[28] G. Boole. *An Investigation of the Laws of Thought on Which are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, London, 1854.

[29] G. E. P. Box. Non-normality and tests on variances. *Biometrika*, 40:318–335, 1953.

[30] G. E. P. Box and N. R. Draper. *Empirical model-building and response surfaces*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons Inc., New York, 1987.

[31] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[32] H. Brandsma. Initial topologies and embeddings. *Topology Q&A Board*, 2011. http://at.yorku.ca/cgi-bin/bbqa?forum=ask_a_topologist;task=show_msg;msg=2719.0002.

[33] L. Breiman, L. Le Cam, and L. Schwartz. Consistent estimates and zero-one sets. *Ann. Math. Statist.*, 35:157–161, 1964.

[34] C. Castaing and M. Valadier. *Convex Analysis and Measurable Multifunctions*. Lecture Notes in Mathematics, Vol. 580. Springer-Verlag, Berlin, 1977.

[35] B. Clarke. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *J. Mach. Learn. Res.*, 4(4):683–712, 2004.

[36] C. Costantini, S. Levi, and J. Pelant. Infima of hyperspace topologies. *Mathematika*, 42(1):67–86, 1995. http://dx.doi.org/10.1112/S0025579300011360.

[37] D. D. Cox. An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.*, 21(2):903–923, 1993.

[38] D. J. Daley and D. Vere-Jones. *An Introduction to the Theory of Point Processes. Vol. II*. Probability and its Applications (New York). Springer, New York, second edition, 2008. General theory and structure.

[39] G. Del Maso. *Introduction to* Γ*-convergence*. Birkhäuser, Boston, 1993.

[40] I. Del Prete and M. B. Lignola. On convergence of closed-valued multifunctions. *Boll. Un. Mat. Ital. B (6)*, 2(3):819–834, 1983.

[41] C. Dellacherie and P.-A. Meyer. *Probabilités et Potentiel*. Hermann, Paris, 1975. Chapitres I à IV, Édition entièrement refondue, Publications de l'Institut de Mathématique de l'Université de Strasbourg, No. XV, Actualités Scientifiques et Industrielles, No. 1372.

[42] P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[43] P. W. Diaconis and D. Freedman. Consistency of Bayes estimates for nonparametric regression: normal theory. *Bernoulli*, 4(4):411–444, 1998.

[44] J. L. Doob. Application of the theory of martingales. In *Le Calcul des Probabilités et ses Applications*, Colloques Internationaux du Centre National de la Recherche Scientifique, no. 13, pages 23–27. Centre National de la Recherche Scientifique, Paris, 1949.

[45] J. L. Doob. *Measure Theory*, volume 143 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.

[46] D. Draper. Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. Ser. B*, 57(1):45–97, 1995. With discussion and a reply by the author.

[47] D. Draper. Bayesian model specification: heuristics and examples. In P. Damien, P. Dellaportas, N. G. Polson, and D. A. Stephens, editors, *Bayesian Theory and Applications*. Oxford University Press, 2013.

[48] R. M. Dudley. *Real Analysis and Probability*, volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2002. Revised reprint of the 1989 original.

[49] E. G. Effros. Convergence of closed subsets in a topological space. *Proc. Amer. Math. Soc.*, 16:929–931, 1965.

[50] England and Wales Court of Appeal (Civil Division). Nulty & Ors v. Milton Keynes Borough Council, 2013. [2013] EWCA Civ 15, Case No. A1/2012/0459. http://www.bailii.org/ew/cases/EWCA/Civ/2013/15.html.

[51] J. M. G. Fell. A Hausdorff topology for the closed subsets of a locally compact non-Haussdorff space. *Proc. Amer. Math. Soc.*, 13(3):472–476, 1962.

[52] R. Fortet. Espaces á noyau reproduisant et lois de probabilités des fonctions aléatoires. *Annales de l'I. H. P., Sec. B*, 9(1):41–58, 1973.

[53] D. Freedman. On the Bernstein-von Mises theorem with infinite-dimensional parameters. *Ann. Statist.*, 27(4):1119–1140, 1999.

[54] D. A. Freedman. On the asymptotic behavior of Bayes' estimates in the discrete case. *Ann. Math. Statist.*, 34:1386–1403, 1963.

[55] D. A. Freedman. On the asymptotic behavior of Bayes estimates in the discrete case. II. *Ann. Math. Statist.*, 36:454–456, 1965.

[56] Z. Frolik. On bianalytic spaces. *Czechoslovak Math. J.*, 13 (88):561–573, 1963.

[57] Z. Frolik. Absolute Borel and Souslin sets. *Pacific J. Math.*, 32:663–683, 1970.

[58] T. Fushiki. Bootstrap prediction and Bayesian prediction under misspecified models. *Bernoulli*, 11(4):747–758, 2005. http://dx.doi.org/10.3150/bj/1126126768.

[59] A. Gelman. Objections to Bayesian statistics. *Bayesian Anal.*, 3(3):445–449, 2008.

[60] S. Ghosal. The Dirichlet process, related priors and posterior asymptotics. In *Bayesian Nonparametrics*, Camb. Ser. Stat. Probab. Math., pages 35–79. Cambridge Univ. Press, Cambridge, 2010.

[61] P. Grünwald and J. Langford. Suboptimal behavior of Bayes and MDL in classification under misspecification. In *Learning theory*, volume 3120 of *Lecture Notes in Comput. Sci.*, pages 331–347. Springer, Berlin, 2004.

[62] Paul Gustafson. On measuring sensitivity to parametric model misspecification. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(1):81–94, 2001.

[63] Jerry A. Hausman and William E. Taylor. A generalized specification test. *Econom. Lett.*, 8(3):239–245, 1981. http://dx.doi.org/10.1016/0165-1765(81)90073-2.

[64] C. Hess. Loi de probabilité des ensembles aléatoires á valeurs fermées dans un espace métrique separable. *C. R. Acad. Sci. Paris, Series I*, 296:883–886, 1983.

[65] C. Hess. Contributions á l'étude de la mesurabilité, de la loi de probabilité, et de la convergence des multifunctions. *Thése d'état, Montpellier*, 1986.

[66] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35:73–101, 1964.

[67] P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proc. Fifth Berkeley Sympos. Math. Statist. and Probability (Berkeley, Calif., 1965/66), Vol. I: Statistics*, pages 221–233. Univ. California Press, Berkeley, Calif., 1967.

[68] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. Wiley Series in Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second edition, 2009. http://dx.doi.org/10.1002/9780470434697.

[69] A. Jakubowski. The Skorokhod space in functional convergence: a short introduction. *Skorokhod Space. 50 Years On*, 17-23 June, Kiev, Ukraine, 2007. http://www-users.mat.uni.torun.pl/~adjakubo/kievtopologies.pdf.

[70] O. Kallenberg. *Random Measures*. Akademie-Verlag, Berlin, 1975. Schriftenreihe des Zentralinstituts für Mathematik und Mechanik bei der Akademie der Wissenschaften der DDR, Heft 23.

[71] A. S. Kechris. *Classical Descriptive Set Theory*. Graduate Texts in Mathematics. Springer-Verlag, New York, 1995.

[72] J. M. Keynes. *A Treatise on Probability*. Macmillan and Co., London, 1921.

[73] B. J. K. Kleijn and A. W. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. *Ann. Statist.*, 34(2):837–877, 2006. http://dx.doi.org/10.1214/009053606000000029.

[74] B. J. K. Kleijn and A. W. van der Vaart. The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat.*, 6:354–381, 2012. http://dx.doi.org/10.1214/12-EJS675.

[75] V. P. Kuznetsov. *Intervalnye Statisticheskie Modeli [Interval Statistical Models]*. "Radio i Svyaz′", Moscow, 1991.

[76] L. Le Cam. On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. California Publ. Statist.*, 1:277–329, 1953.

[77] H. Lian. On rates of convergence for posterior distributions under misspecification. *Comm. Statist. Theory Methods*, 38(11-12):1893–1900, 2009.

[78] D. Malakoff. Bayes offers a 'new' way to make sense of numbers. *Science*, 286(5444):1460–1464, 1999. http://dx.doi.org/10.1126/science.286.5444.1460.

[79] R. Martin and L. Hong. On convergence rates of Bayesian predictive densities and posterior distributions. *arXiv*, 1210.0103v1, 2012.

[80] G. Matheron. *Random Sets and Integral Geometry*. Wiley, New York, 1975.

[81] S. B. McGrayne. *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy*. Yale University Press, 2012.

[82] E. Michael. Topologies on spaces of subsets. *Trans. Amer. Math. Soc.*, 71:152–182, 1951.

[83] I. Molchanov. *Theory of Random Sets*. Probability and its Applications (New York). Springer-Verlag London Ltd., London, 2005.

[84] R. Nickl. Statistical Theory, 2012. http://www.statslab.cam.ac.uk/~nickl/Site/__files/stat.pdf.

[85] H. Owhadi and C. Scovel. Brittleness of Bayesian inference and new Selberg formulas. 2013.

[86] H. Owhadi, C. Scovel, T. J. Sullivan, M. McKerns, and M. Ortiz. Optimal Uncertainty Quantification. *SIAM Review*, To appear, 2013. Preprint at arXiv:1009.0679.

[87] J. C. Oxtoby. *Measure and Category. A Survey of the Analogies Between Topological and Measure Spaces*. Springer-Verlag, New York, 1971. Graduate Texts in Mathematics, Vol. 2.

[88] Grünwald P. Bayesian inconsistency under misspecification, 2006.

[89] R. R. Phelps. *Lectures on Choquet's Theorem*, volume 1757 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, second edition, 2001. http://dx.doi.org/10.1007/b76887.

[90] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1998. http://dx.doi.org/10.1007/978-3-642-02431-3.

[91] H. Rosenthal. Differences of bounded semi-continuous functions, I. 1994. http://arxiv.org/pdf/math.FA/9406217.

[92] H. Rosenthal. personal communication. 2011.

[93] B. Rustem, R. G. Becker, and W. Marty. Robust min-max portfolio strategies for rival forecast and risk scenarios. *J. Econom. Dynam. Control*, 24(11-12):1591–1621, 2000. Computational aspects of complex securities.

[94] M.-F. Sainte-Beuve. On the extension of von Neumann-Aumann's theorem. *J. Functional Analysis*, 17:112–129, 1974.

[95] F. J. Samaniego. *A comparison of the Bayesian and frequentist approaches to estimation*. Springer Series in Statistics. Springer, New York, 2010. http://dx.doi.org/10.1007/978-1-4419-5941-6.

[96] L. Schwartz. On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 4:10–26, 1965.

[97] L. Schwartz. *Radon Measures on Arbitrary Topological Spaces and Cylindrical Measures*. Oxford Univ. Press, Oxford, 1974.

[98] J. E. Smith. Generalized Chebychev inequalities: theory and applications in decision analysis. *Oper. Res.*, 43(5):807–825, 1995. http://dx.doi.org/10.1287/opre.43.5.807.

[99] E. H. Spanier. *Algebraic Topology*. Springer-Verlag, New York, 1966.

[100] I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2(1):67–93, 2002. http://dx.doi.org/10.1162/153244302760185252.

[101] I. Steinwart and A. Christmann. *Support Vector Machines*. Information Science and Statistics. Springer, New York, 2008.

[102] I. Steinwart, D. Hush, and C. Scovel. An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Trans. Inform. Theory*, 52(10):4635–4643, 2006. http://dx.doi.org/10.1109/TIT.2006.881713.

[103] I. Steinwart, D. Hush, and C. Scovel. Function classes that approximate the Bayes risk. 4005:79–93, 2006. http://dx.doi.org/10.1007/11776420_9.

[104] I. Steinwart and C. Scovel. Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.*, 35(2):575–607, 2007. http://dx.doi.org/10.1214/009053606000001226.

[105] I. Steinwart and C. Scovel. Mercer's theorem on general domains: on the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35(3):363–417, 2012. http://dx.doi.org/10.1007/s00365-012-9153-3.

[106] A. H. Stone. Non-separable Borel sets. In *General Topology and its Relations to Modern Analysis and Algebra (Proc. Sympos. Prague, 1961)*, pages 341–342. Academic Press, New York, 1962.

[107] R. V. Telgársky. Topological games: on the 50th anniversary of the Banach–Mazur game. *Rocky Mountain J. Math.*, 17(2):227–276, 1987. http://dx.doi.org/10.1216/RMJ-1987-17-2-227.

[108] F. Topsøe. *Topology and Measure*. Lecture Notes in Mathematics, Vol. 133. Springer-Verlag, Berlin, 1970.

[109] H. Tuy. *Convex Analysis and Global Optimization*, volume 22 of *Nonconvex Optimization and its Applications*. Kluwer Academic Publishers, Dordrecht, 1998.

[110] R. von Mises. *Mathematical Theory of Probability and Statistics*. Edited and Complemented by Hilda Geiringer. Academic Press, New York, 1964.

[111] H. von Weizsäcker and G. Winkler. Integral representation in the set of solutions of a generalized moment problem. *Math. Ann.*, 246(1):23–32, 1979/80. http://dx.doi.org/10.1007/BF01352023.

[112] S. Walker. New approaches to Bayesian consistency. *Ann. Statist.*, 32(5):2028–2043, 2004. http://dx.doi.org/10.1214/009053604000000409.

[113] S. Walker and N. L. Hjort. On Bayesian consistency. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 63(4):811–821, 2001.

[114] P. Walley. *Statistical Reasoning with Imprecise Probabilities*, volume 42 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London, 1991.

[115] L. Wasserman. Asymptotic properties of nonparametric Bayesian procedures. In *Practical nonparametric and semiparametric Bayesian statistics*, volume 133 of *Lecture Notes in Statist.*, pages 293–304. Springer, New York, 1998.

[116] L. Wasserman, M. Lavine, and R. L. Wolpert. Linearization of Bayesian robustness problems. *J. Statist. Plann. Inference*, 37(3):307–316, 1993. http://dx.doi.org/10.1016/0378-3758(93)90109-J.

[117] L. A. Wasserman. Prior envelopes based on belief functions. *Ann. Statist.*, 18(1):454–464, 1990. http://dx.doi.org/10.1214/aos/1176347511.

[118] K. Weichselberger. The theory of interval-probability as a unifying concept for uncertainty. *Internat. J. Approx. Reason.*, 24(2-3):149–170, 2000. Reasoning with imprecise probabilities (Ghent, 1999).

[119] H. White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

[120] R. A. Wijsman. Convergence of sequences of convex sets, cones and functions. II. *Trans. Amer. Math. Soc.*, 123:32–45, 1966.

[121] G. Winkler. Extreme points of moment sets. *Math. Oper. Res.*, 13(4):581–587, 1988. http://dx.doi.org/10.1287/moor.13.4.581.

[122] H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *J. Mach. Learn. Res.*, 10:2741–2775, 2009. http://dx.doi.org/10.1109/IJCNN.2009.5179093.

[123] L. Zhu, T. F. Coleman, and Y. Li. Min-max robust and cvar robust mean-variance portfolios. *Journal of Risk*, 11(3):1–31, 2009.

[124] L. Zsilinszky. Topological games and hyperspace topologies. *Set-Valued Anal.*, 6(2):187–207, 1998. http://dx.doi.org/10.1023/A:1008669420995.